



Volume 106  
Number 3

August 2014

Published quarterly  
by the  
American Psychological  
Association

ISSN 0022-0663

# Journal of Educational Psychology

Arthur C. Graesser, *Editor*

Jill Fitzgerald, *Associate Editor*

David Francis, *Associate Editor*

Susan Goldman, *Associate Editor*

Young-Suk Kim, *Associate Editor*

Robert Klassen, *Associate Editor*

David N. Rapp, *Associate Editor*

Susan Sonnenschein, *Associate Editor*

Birgit Spinath, *Associate Editor*

Roman Taraban, *Associate Editor*

Jennifer Wiley, *Associate Editor*

Christopher A. Wolters, *Associate Editor*

[www.apa.org/pubs/journals/edu](http://www.apa.org/pubs/journals/edu)

**Marygrove College Library**  
**8425 West McNichols Road**  
**Detroit, MI 48221**



## Editor

Arthur C. Graesser, *University of Memphis*

## Associate Editors

Jill Fitzgerald, *University of North Carolina at Chapel Hill, Emeritus*  
David Francis, *University of Houston*  
Susan Goldman, *University of Illinois, Chicago*  
Young-Suk Kim, *Florida State University*  
Robert Klassen, *The University of York, United Kingdom*  
David N. Rapp, *Northwestern University*  
Susan Sonnenschein, *University of Maryland*  
Birgit Spinath, *University of Heidelberg, Heidelberg, Germany*  
Roman Taraban, *Texas Tech University*  
Jennifer Wiley, *University of Illinois at Chicago*  
Christopher A. Wolters, *The Ohio State University*

## Incoming Editor

Steve Graham, *Arizona State University*

For a complete list of Incoming Associate Editors and Consulting Editors see  
<http://www.apa.org/pubs/journals/edu/>

## Chief Editorial Assistant

Jean Edgar, *University of Memphis*

## Advisory Editors

Mary D. Ainley, *University of Melbourne, Australia*  
Vincent Alevén, *Carnegie Mellon University*  
Patricia Alexander, *University of Maryland, College Park*  
Richard L. Allington, *University of Tennessee*  
Ellen R. Altermatt, *Hanover College*  
Ivan Ash, *Old Dominion University*  
Carole Beal, *University of Arizona*  
David A. Bergin, *University of Missouri, Columbia*  
Daniel Bolt, *University of Wisconsin, Madison*  
Mimi Bong, *Ewha Womans University, Seoul, Korea*  
Julie L. Booth, *Temple University*  
Jason Braasch, *University of Memphis*  
M. Anne Britt, *Northern Illinois University*  
Scott Brown, *University of Connecticut*  
Eric S. Buhs, *University of Nebraska, Lincoln*  
Adriana G. Bus, *Leiden University, The Netherlands*  
Kirsten R. Butcher, *University of Utah*  
Robert Calfee, *University of California, Riverside*  
Martha Carr, *University of Georgia*  
Kwansu Cho, *University of Missouri, Columbia*  
Timothy Cleary, *University of Wisconsin, Milwaukee*  
Anne E. Cook, *University of Utah*  
Kai Cortina, *University of Michigan*  
Jennifer Cromley, *Temple University*  
H. Michael Crowson, *University of Oklahoma*  
Anne E. Cunningham, *University of California, Berkeley*  
Teresa K. DeBacker, *The University of Oklahoma*  
Sidney D'Mello, *University of Notre Dame*  
John Dunlosky, *Kent State University*  
Amanda M. Durik, *Northern Illinois University*  
Gary Feng, *Educational Testing Service*  
J. D. Fletcher, *Institute for Defense Analyses*  
Lynn S. Fuchs, *Vanderbilt University*  
Linda Gambrell, *Clemson University*  
James P. Gee, *Arizona State University*  
Arthur M. Glenberg, *Arizona State University*  
Adele E. Gottfried, *California State University*  
Steve Graham, *Arizona State University*  
Barbara A. Greene, *University of Oklahoma*  
John Guthrie, *University of Maryland*  
Douglas Hacker, *University of Utah*  
Vernon C. Hall, *Syracuse University*  
Jill Hamm, *University of North Carolina, Chapel Hill*  
John Hattie, *University of Auckland, New Zealand*  
Mary Hegarty, *University of California, Santa Barbara*  
Flaviu A. Hodis, *Victoria University of Wellington, New Zealand*  
Jan N. Hughes, *Texas A&M University*  
Slava Kalyuga, *University of South Wales, Australia*  
Avi Kaplan, *Temple University*  
Beth Kurtz-Costes, *University of North Carolina, Chapel Hill*  
Dan Lapsley, *University of Notre Dame*  
Willy Lens, *University of Leuven, Belgium*  
Elizabeth A. Linnenbrink-Garcia, *Duke University*  
Robert Lorch, *University of Kentucky*  
Joseph P. Magliano, *Northern Illinois University*  
Andrew Martin, *University of Sydney, Australia*  
Andrew J. Mashburn, *Portland State University*  
Linda Mason, *Pennsylvania State University*  
Richard E. Mayer, *University of California, Santa Barbara*  
Charles MacArthur, *University of Delaware*  
Catherine McBride-Chang, *The Chinese University of Hong Kong, China*  
Nicole M. McNeil, *University of Notre Dame*  
Debra K. Meyer, *Elmhurst College*  
Keith Millis, *Northern Illinois University*  
Alexandre J. S. Morin, *University of Western Sydney, Australia*  
Tamera B. Murdock, *University of Missouri, Kansas City*  
P. Karen Murphy, *Pennsylvania State University*  
Mitchell J. Nathan, *University of Wisconsin, Madison*  
Nikolaso Ntoumanis, *University of Birmingham, United Kingdom*  
E. Michael Nussbaum, *University of Nevada, Las Vegas*  
Rollanda E. O'Connor, *University of California, Riverside*  
Harry O'Neil, *University of Southern California*  
Tenaha O'Reilly, *Educational Testing Service*

Philip Parker, *University of Western Sydney, Australia*  
Helen Patrick, *Purdue University*  
Erika Patall, *University of Texas, Austin*  
Reinhard Pekrun, *University of Munich, Germany*  
Yaacov Petscher, *Florida State University*  
Gary Phye, *Iowa State University*  
Keenan Pituch, *University of Texas, Austin*  
Jan L. Plass, *New York University*  
Katherine Rawson, *Kent State University*  
Robert Renaud, *University of Manitoba, Canada*  
Alexander Renkl, *University of Freiburg, Germany*  
Catherine Richards-Tutor, *California State University, Long Beach*  
Bethany Rittle-Johnson, *Vanderbilt University*  
Daniel Robinson, *University of Texas, Austin*  
Philip Rodkin, *University of Illinois at Urbana-Champaign*  
Christine M. Rubie-Davies, *University of Auckland, New Zealand*  
Christopher A. Sanchez, *Arizona State University*  
Katherine Scheiter, *Knowledge Media Research Center, Germany*  
Marlene Schommer-Aikins, *Wichita State University*  
Gregory Schraw, *University of Nevada, Las Vegas*  
Dale Schunk, *University of North Carolina, Greensboro*  
Christian D. Schunn, *University of Pittsburgh*  
Paula J. Schwanenflugel, *University of Georgia*  
Colleen M. Seifert, *University of Michigan*  
Timothy Shanahan, *University of Illinois, Chicago*  
Gale M. Sinatra, *University of Southern California*  
Einar M. Skaalvik, *Norwegian University of Science and Technology, Norway*  
John Sweller, *University of New South Wales, Australia*  
Keith Thiede, *Boise State University*  
Theresa A. Thorkildsen, *University of Illinois, Chicago*  
Wendy Troop-Gordon, *North Dakota State University*  
Chia-Wen Tsai, *Ming Chuan University-Taiwan*  
Timothy Urdan, *Santa Clara University*  
Ellen Usher, *University of Kentucky*  
Regina Vollmeyer, *University of Frankfurt, Germany*  
Jeffrey Walczyk, *Louisiana Technical University*  
Charles A. Weaver III, *Baylor University*  
Joanna P. Williams, *Columbia University*  
Phil Winne, *Simon Fraser University, Canada*  
Moshe M. Zeidner, *University of Haifa, Israel*

The main purpose of the *Journal of Educational Psychology*® is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

**Single Issues, Back Issues, and Back Volumes:** For information regarding single issues, back issues, or back volumes, write to Order Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242; call 202-336-5600 or 800-374-2721; or visit [www.apa.org/pubs/journals/subscriptions.aspx](http://www.apa.org/pubs/journals/subscriptions.aspx)

**Manuscripts:** Submit manuscripts electronically through the Manuscript Submissions Portal found at [www.apa.org/pubs/journals/edu](http://www.apa.org/pubs/journals/edu) according to the Instructions to Authors found elsewhere in this issue (see table of contents). Correspondence regarding manuscripts should be sent to the Editor, Art Graesser, Journal of Educational Psychology, 202 Psychology Building University of Memphis, Memphis, TN 38152-3230. The opinions and statements published are the responsibility of the authors, and such opinions and statements do not necessarily represent the policies of APA or the views of the Editor.

**Copyright and Permission:** Those who wish to reuse APA-copyrighted material in a non-APA publication must secure from APA written permission to reproduce a journal article in full or journal text of more than 800 cumulative words or more than 3 tables and/or figures. APA normally grants permission contingent on permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of a fee of \$25 per page. Libraries are permitted to photocopy beyond the limits of the U.S. copyright law: (1) post-1977 articles, provided the per-copy fee in the code for this journal (0022-0663/14/\$12.00) is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923; (2) pre-1978 articles, provided that the per-copy fee stated in the Publishers' Fee List is paid through the Copyright Clearance Center. For more information along with a permission form, go to [www.apa.org/about/contact/copyright/index.aspx](http://www.apa.org/about/contact/copyright/index.aspx)

**Electronic Access:** APA members who subscribe to this journal have automatic access to all issues of the journal in the PsycARTICLES® full-text database. See <http://my.apa.org/access.html>.

**Reprints:** Authors may order reprints of their articles from the printer when they receive proofs.

**APA Journal Staff:** Susan J. A. Harris, *Senior Director, Journals Program*; John Breithaupt, *Director, Journal Production Services*; Stephanie Pollock, *Account Manager*; Jodi Ashcraft, *Director, Advertising Sales and Exhibits*.

**Journal of Educational Psychology**® (ISSN 0022-0663) is published quarterly (February, May, August, November) in one volume per year by the American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Subscriptions are available on a calendar year basis only (January through December). The 2014 rates follow: *Nonmember Individual*: \$208 Domestic, \$237 Foreign, \$250 Air Mail. *Institutional*: \$751 Domestic, \$800 Foreign, \$815 Air Mail. *APA Member*: \$89. *APA Student Affiliate*: \$62. Write to Subscriptions Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Printed in the U.S.A. Periodicals postage paid at Washington, DC, and at additional mailing offices. POSTMASTER: Send address changes to *Journal of Educational Psychology*, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

Effective with the 1986 volume, this journal is printed on acid-free paper.

*Journal of Educational Psychology*® is a registered trademark of the American Psychological Association.



**Special Section: Computer-Based Assessment of Cross-Curricular Skills and Processes**

**Guest Editors: Samuel Greiff, Romain Martin, and Birgit Spinath**

- |     |  |
|-----|--|
| 605 | Introduction to the Special Section on Computer-Based Assessment of Cross-Curricular Skills and Processes<br><i>Samuel Greiff, Romain Martin, and Birgit Spinath</i>   |
| 608 | The Time on Task Effect in Reading and Problem Solving Is Moderated by Task Difficulty and Skill: Insights From a Computer-Based Large-Scale Assessment<br><i>Frank Goldhammer, Johannes Naumann, Annette Stelter, Krisztina Tóth, Heiko Rölke, and Eckhard Klieme</i> |
| 627 | The Role of Time on Task in Computer-Based Low-Stakes Assessment of Cross-Curricular Skills<br><i>Sirkku Kupiainen, Mari-Pauliina Vainikainen, Jukka Marjanen, and Jarkko Hautamäki</i>  |
| 639 | Computer-Based Assessment of School Readiness and Early Reasoning<br><i>Benő Csapó, Gyöngyvér Molnár, and József Nagy</i>  |
| 651 | Toward Automated Computer-Based Visualization and Assessment of Team-Based Performance<br><i>Dirk Ifenthaler</i>   |
| 666 | The Computer-Based Assessment of Complex Problem Solving and How It Is Influenced by Students' Information and Communication Technology Literacy<br><i>Samuel Greiff, André Kretzschmar, Jonas C. Müller, Birgit Spinath, and Romain Martin</i>                        |
| 681 | Differential Relations Between Facets of Complex Problem Solving and Students' Immigration Background<br><i>Philipp Sonleitner, Martin Brunner, Ulrich Keller, and Romain Martin</i>   |

## Articles

- |     |  |
|-----|--|
| 696 | Boredom and Academic Achievement: Testing a Model of Reciprocal Causation<br><i>Reinhard Pekrun, Nathan C. Hall, Thomas Goetz, and Raymond P. Perry</i>                                |
| 711 | Perfectionism and Motivation of Adolescents in Academic Contexts<br><i>Mimi Bong, Arum Hwang, Arum Noh, and Sung-il Kim</i>  |
| 730 | The Contribution of Adolescent Effortful Control to Early Adult Educational Attainment<br><i>Marie-Hélène Véronneau, Kristina Hiatt Racer, Gregory M. Fosco, and Thomas J. Dishion</i> |



- 744 Academic Self-Handicapping and Achievement: A Meta-Analysis  
*Malte Schwinger, Linda Wirthwein, Gunnar Lemmer, and Ricarda Steinmayr*
- 762 Capturing the Complexity: Content, Type, and Amount of Instruction and Quality of the Classroom Learning Environment Synergistically Predict Third Graders' Vocabulary and Reading Comprehension Outcomes  
*Carol McDonald Connor, Mercedes Spencer, Stephanie L. Day, Sarah Giuliani, Sarah W. Ingebrand, Leigh McLean, and Frederick J. Morrison*
- 779 Text Comprehension Mediates Morphological Awareness, Syntactic Processing, and Working Memory in Predicting Chinese Written Composition Performance  
*Connie Qun Guan, Feifei Ye, Richard K. Wagner, Wanjin Meng, and Che Kan Leong*
- 799 Impact of a Teacher-Led Intervention on Preference for Self-Regulated Learning, Finding Main Ideas in Expository Texts, and Reading Comprehension  
*Heidrun Stoeger, Christine Sontag, and Albert Ziegler*
- 815 Can Babies Learn to Read? A Randomized Trial of Baby Media  
*Susan B. Neuman, Tanya Kaefer, Ashley Pinkham, and Gabrielle Strouse*
- 831 Does Cognitive Strategy Training on Word Problems Compensate for Working Memory Capacity in Children With Math Difficulties?  
*H. Lee Swanson*
- 849 Learning With Retrieval-Based Concept Mapping  
*Janell R. Blunt and Jeffrey D. Karpicke*
- 859 Can Parents' Involvement in Children's Education Offset the Effects of Early Insensitivity on Academic Functioning?  
*Jennifer D. Monti, Eva M. Pomerantz, and Glenn I. Roisman*
- 870 Strengthening Bullying Prevention Through School Staff Connectedness  
*Lindsey M. O'Brennan, Tracy E. Waasdorp, and Catherine P. Bradshaw*
- 881 Testing the Theory of Successful Intelligence in Teaching Grade 4 Language Arts, Mathematics, and Science  
*Robert J. Sternberg, Linda Jarvin, Damian P. Birney, Adam Naples, Steven E. Stemler, Tina Newman, Renate Otterbach, Carolyn Parish, Judy Randi, and Elena L. Grigorenko*

---

## Other

- 743 E-Mail Notification of Your Latest Issue Online!
- 900 Instructions to Authors
- 680 Subscription Order Form



# Introduction to the Special Section on Computer-Based Assessment of Cross-Curricular Skills and Processes

Samuel Greiff and Romain Martin  
University of Luxembourg

Birgit Spinath  
Heidelberg University

*Keywords:* computer-based assessment, cross-curricular skills, behavioral processes, domain-general

This special section presents a collection of articles that were submitted to *Journal of Educational Psychology* in response to a call for papers on computer-based assessment of cross-curricular skills and processes. The development of innovative computer-based assessment instruments that target cross-curricular skills and processes and the validation of these instruments within educational psychology has been a field of ongoing scientific inquiry with substantial research activity in recent years. After a selective and stringent peer-review process, this special section includes six articles that report cutting-edge research and present a cross-section of different topics.

Why is a special section on computer-based assessment of cross-curricular skills and processes both timely and important to researchers interested in the field of educational psychology? There are a number of good reasons, but a major reason is that the cognitive and interpersonal skills necessary for successful participation in society have undergone great changes in recent decades. Studies have shown that tasks at school, university, and work have become more demanding and less bound to single-subject matters or domains (e.g., Autor, Levy, & Murnane, 2003; Spitz-Oener, 2006). Tasks now more often involve cross-curricular, nonroutine, and complex skills and processes (e.g., problem solving) that are applicable in diverse situations and content areas (Greiff et al., 2013; Hautamäki et al., 2002). Mayer and Wittrock (2006) highlighted the importance of problem solving as one prime example of a cross-curricular skill for educational psychologists. In fact, they proposed that helping students become better problem solvers is one of the greatest challenges in educational psychology.

Consequently, the development of assessment instruments to measure cross-curricular skills and processes as well as their validation has been an ongoing field of inquiry in psychometrics and educational psychology alike. However, the dynamic and interactive nature of these skills implies that their assessment may not lie within reach of classical paper-and-pencil instruments.

Fortunately, the advent of computers in virtually any setting of educational assessment has allowed for the emergence of innovative assessment procedures. In addition to offering increased flexibility, computer-administered tests record log-file data during task execution, thus providing further insight into behavioral processes that are not captured by final performance data. For instance, time on task may be used to better understand how students become involved in a proposed task or to yield information about the type and quality of cognitive processing that occurs while students work on an educational assessment.

From an applied perspective, computer-based assessment environments and the use of computer-generated log-file data are now found in a large range of educational settings, including international large-scale assessments such as the Programme for International Student Assessment (PISA) and the Programme for the International Assessment of Adult Competencies (PIAAC). Cross-curricular skills have become integral parts of the assessment framework in interactive problem solving (Organisation for Economic Co-operation and Development [OECD], 2010), collaborative problem solving (OECD, 2012), problem solving in technology-rich environments (OECD, 2009), and electronic reading assessment (OECD, 2011). In addition, process data are now implemented in the scoring procedures of large-scale educational assessments, for example, to correct for obvious guessing as identified through a lack of the required exploration behavior or to integrate behavioral data as potential performance indicators that go beyond merely scoring the number of correct answers.

As a consequence, research concerning the setup and use of computer-based assessment instruments in educational contexts is quickly emerging and has great relevance to researchers and practitioners alike. This special section in *Journal of Educational Psychology* pays tribute to the general need for rigorous empirical research in this field. This need is illustrated through the assessment of cross-curricular skills in particular, stressing the importance of developing a theoretical understanding of these skills and the added value of computerized assessments gained through the setup of interactive and complex assessment environments and the use of log-file data. This special section is composed of articles that report on the development of theoretically sound and scientifically validated assessment instruments for cross-curricular skills and on the benefits of methodological advances associated with computer-based assessment, such as the benefits of log-file analyses for assessing classical and cross-curricular cognitive abilities. Articles are related to assessment issues in education, and some of them exhibit strong ties to international or national large-

---

Samuel Greiff and Romain Martin, Department of Psychology, University of Luxembourg, Luxembourg-Kirchberg, Luxembourg; Birgit Spinath, Department of Psychology, Heidelberg University, Heidelberg, Germany.

This research was funded by a grant from the Fonds National de la Recherche Luxembourg (ATTRACT “ASKI21”).

Correspondence concerning this article should be addressed to Samuel Greiff, Department of Psychology, University of Luxembourg, 6, rue Richard Coudenhove Kalergi, 1359 Luxembourg-Kirchberg, Luxembourg. E-mail: samuel.greiff@uni.lu



scale efforts. For example, some contributions uncover behavioral interaction patterns that are not reflected in final performance data and relate these patterns to psychological theories and relevant educational outcomes within a large-scale assessment. Thus, the common denominator of all articles published in the special section is that they exploit the computer in an innovative manner and substantially widen the scope of our view on students' skills. Much of the research in this special section is embedded in the context of large-scale educational assessments, but some of it is experimental or draws on selective subgroups of student populations.

In the first contribution (Goldhammer, Naumann, Stelter, Tóth, Rölke, & Klieme, 2014), the authors provided insights into behavioral processes that, until recently, were exclusively addressed in experimental settings. They elaborated on differential effects of the meaning of time on task in problem solving and in reading based on a representative German sample from the PIAAC field trial data. In a related way, the second contribution (Kupiainen, Vainikainen, Marjanen, & Hautamäki, 2014) investigated the role of time on task as an indicator of students' investment and the subsequent effect on students' achievement. Both articles make use of log files and show how this approach not only broadens the understanding of assessment but also advances theory in the field of educational psychology and may ultimately lead to the development of interventions that can be used in the classroom.

The third contribution (Csapó, Molnár, & Nagy, 2014) demonstrates that psychometric properties can be optimized through computer-based test delivery even in an assessment of school readiness at a very young age. This illustrates that computers as assessment instruments can be used across a variety of age groups—a topic with very limited information until now. The fourth contribution (Ifenthaler, 2014) shows how computers can be used not only to collect large amounts of data but also to process and to automatically score the data in the context of team-based processes and performance. By doing so, Ifenthaler investigated team effectiveness, an area that is very relevant to the assessment of cooperation and collaboration in large-scale educational assessments. Collaborative problem solving will be assessed in the PISA 2015 cycle.

The fifth contribution (Greiff, Kretzschmar, Müller, Spinath, & Martin, 2014) addressed complex problem solving, a phenomenon particularly relevant to cross-curricular skills. The authors related complex problem solving to intelligence and computer skills and showed in three different samples that the added value of complex problem solving cannot be traced back to an indirect assessment of computer skills. In fact, the added value seems to originate from complex cognitive processes associated with computer-simulated problem solving tasks. Further investigating the skill of complex problem solving, the sixth contribution (Sonnleitner, Brunner, Keller, & Martin, 2014) reported that computer-based simulations of complex cognitive processes may be less influenced than paper-and-pencil tests of intelligence by students' cultural backgrounds, thus yielding a less biased and fairer assessment of cognitive skills for disadvantaged groups or minorities. These two articles provide strong empirical support for the claim that computer-based assessment allows psychologists to widen their scope to new cognitive constructs that cannot be accessed via classical paper-and-pencil-based measures.

These six contributions in the special section span a wide array of different topics. They do not cover all topics relevant in the field, but they do cover important topics for scientists interested in

new developments in educational psychology. All articles in this special section were subjected to the normal rigorous process of anonymous peer review. Articles in which one of the guest editors was involved as a contributing author were not reviewed or edited by any of the other guest editors. Editorship and authorship were strictly separated in accordance with American Psychological Association guidelines.

Over 20 years ago, Bunderson, Inouye, and Olsen (1989) predicted that new generations of computer-based assessment instruments would swiftly evolve along with a rapid decline in paper-and-pencil testing. From today's perspective, the shift toward a new generation of tests that allow for the assessment of more general and transversal skills and that exploit process data as a standard procedure has been much slower than was anticipated in the late 1980s. Considering how long computers have been available, Williamson, Bejar, and Mislevy (2006) observed that the exploitation of the potential that lay in computer-based assessment has been slower than expected. However, assessment is now at a transition point where many former barriers, such as availability of computer equipment in the classroom or the general level of computer literacy (cf. digital natives; Prensky, 2001), have become less relevant. As a consequence, computer-based assessment is widely available and accepted today, even if its potential for added value still needs to be established and fostered. The contributions in this special section are committed to advancing the knowledge of computer-based assessment in educational contexts. In doing so, our goal of this special section is to enhance the understanding of the assessment of cross-curricular skills and processes at different educational levels and to explain the process of skill acquisition by developing and testing adequate models. We sincerely hope that you enjoy reading this special section in the *Journal of Educational Psychology*.

## References

- Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118, 1279–1333. doi:10.1162/003355303322552801
- Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational measurement* (pp. 367–407). New York, NY: Macmillan.
- Csapó, B., Molnár, G., & Nagy, J. (2014). Computer-based assessment of school readiness and early reasoning. *Journal of Educational Psychology*, 106, 639–650. doi:10.1037/a0035756
- Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time-on-task effect in reading and problem solving is moderated by item difficulty and ability: Insights from computer-based large-scale assessment. *Journal of Educational Psychology*, 106, 608–628. doi:10.1037/a0034716
- Greiff, S., Kretzschmar, A., Müller, J. C., Spinath, B., & Martin, R. (2014). The computer-based assessment of complex problem solving and how it is influenced by students' information and communication technology literacy. *Journal of Educational Psychology*, 106, 666–680. doi:10.1037/a0035426
- Greiff, S., Wüstenberg, S., Molnar, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational settings—something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology*, 105, 364–379. doi:10.1037/a0031856
- Hautamäki, J., Arinen, P., Eronen, S., Hautamäki, A., Kupiainen, S., Lindblom, B., . . . Scheinin, P. (2002). *Assessing learning-to-learn. A framework*. Helsinki, Finland: National Board of Education & Univer-



- sity of Helsinki. Retrieved from [http://www.oph.fi/download/47716\\_learning.pdf](http://www.oph.fi/download/47716_learning.pdf)
- Kupiainen, S., Vainikainen, M.-P., Marjanen, J., & Hautamäki, J. (2014). The role of time on task in computer-based low-stakes assessment of cross-curricular skills. *Journal of Educational Psychology, 106*, 627–638. doi:10.1037/a0035507
- Ifenthaler, D. (2014). Toward automated computer-based visualization and assessment of team-based performance. *Journal of Educational Psychology, 106*, 651–665. doi:10.1037/a0035505
- Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 287–303). Mahwah, NJ: Erlbaum.
- Organisation for Economic Co-operation and Development. (2009). *PIAAC problem solving in technology-rich environments: A conceptual framework (OECD Education Working Papers, No. 36)*. Paris, France: Author.
- Organisation for Economic Co-operation and Development. (2010). *PISA 2012 problem solving framework*. Paris, France: Author.
- Organisation for Economic Co-operation and Development. (2011). *PISA 2009 results: Students on line. Digital technologies and performance*. Paris, France: Author.
- Organisation for Economic Co-operation and Development. (2012, April). *PISA 2015 field trial collaborative problem solving framework*. Paper presented at the 33rd PISA Governing Board meeting, Tallinn, Estonia.
- Prensky, M. (2001). Digital natives, digital immigrants: Part 1. *On the Horizon, 9*, 1–6.
- Sonnleitner, P., Brunner, M., Keller, U., & Martin, R. (2014). Differential relations between facets of complex problems solving and students' immigration background. *Journal of Educational Psychology, 106*, 681–695. doi:10.1037/a0035506
- Spitz-Oener, A. (2006). Technical change, job tasks, and rising educational demands: Looking outside the wage structure. *Journal of Labor Economics, 24*, 235–270. doi:10.1086/499972
- Williamson, D. M., Bejar, I. I., & Mislevy, R. J. (2006). *Automated scoring of complex tasks in computer-based testing: An introduction*. Mahwah, NJ: Erlbaum.

Received November 30, 2013

Revision received November 30, 2013

Accepted December 5, 2013 ■



# The Time on Task Effect in Reading and Problem Solving Is Moderated by Task Difficulty and Skill: Insights From a Computer-Based Large-Scale Assessment

Frank Goldhammer and Johannes Naumann  
German Institute for International Educational Research (DIPF),  
Frankfurt/Main, Germany, and Centre for International Student  
Assessment (ZIB), Frankfurt/Main, Germany

Annette Stelter, Krisztina Tóth, and Heiko Rölke  
German Institute for International Educational Research (DIPF),  
Frankfurt/Main, Germany

Eckhard Klieme  
German Institute for International Educational Research (DIPF), Frankfurt/Main, Germany, and  
Centre for International Student Assessment (ZIB), Frankfurt/Main, Germany

Computer-based assessment can provide new insights into behavioral processes of task completion that cannot be uncovered by paper-based instruments. Time presents a major characteristic of the task completion process. Psychologically, time on task has 2 different interpretations, suggesting opposing associations with task outcome: Spending more time may be positively related to the outcome as the task is completed more carefully. However, the relation may be negative if working more fluently, and thus faster, reflects higher skill level. Using a dual processing theory framework, the present study argues that the validity of each assumption is dependent on the relative degree of controlled versus routine cognitive processing required by a task, as well as a person's acquired skill. A total of 1,020 persons ages 16 to 65 years participated in the German field test of the Programme for the International Assessment of Adult Competencies. Test takers completed computer-based reading and problem solving tasks. As revealed by linear mixed models, in problem solving, which required controlled processing, the time on task effect was positive and increased with task difficulty. In reading tasks, which required more routine processing, the time on task effect was negative and the more negative, the easier a task was. In problem solving, the positive time on task effect decreased with increasing skill level. In reading, the negative time on task effect increased with increasing skill level. These heterogeneous effects suggest that time on task has no uniform interpretation but is a function of task difficulty and individual skill.

**Keywords:** computer-based assessment, time on task, automatic and controlled processing, reading literacy, problem solving

There are two fundamental observations on human performance: the result obtained on a task and the time taken (e.g., Ebel, 1953). In educational assessment, the focus is mainly on the task outcome; behavioral processes that led to the result are usually not considered. One reason may be that traditional assessments are paper-based and, hence, are not suitable for collecting behavioral process data at the task level (cf. Scheuermann & Björnsson, 2009). However, computer-based assessment—besides other advantages, such as increased construct validity (e.g., Sireci & Zenisky, 2006) or improved test design (e.g., van der Linden,

2005)—can provide further insights into the task completion process. This is because in computer-based assessment, log file data can be recorded by the assessment system that allows the researcher to derive theoretically meaningful descriptors of the task completion process. The present study draws on log file data from an international computer-based large-scale assessment to address the question of how time on task is related to the task outcome. As shown in the following, by analyzing the relation of task performance to the time test takers spent on task, we were able to obtain new insights into how the interaction of task and person charac-

---

This article was published Online First February 17, 2014.

Frank Goldhammer and Johannes Naumann, Department of Educational Quality and Evaluation, German Institute for International Educational Research (DIPF), Frankfurt/Main, Germany, and Centre for International Student Assessment (ZIB), Frankfurt/Main, Germany; Annette Stelter, Department of Educational Quality and Evaluation, German Institute for International Educational Research (DIPF), Frankfurt/Main, Germany; Krisztina Tóth and Heiko Rölke, Information Center for Education, German Institute for International Educational Research (DIPF), Frankfurt/Main, Germany; Eckhard Klieme, Department of Educational Quality and Evaluation, German Institute for International

Educational Research (DIPF), Frankfurt/Main, Germany, and Centre for International Student Assessment (ZIB), Frankfurt/Main, Germany.

This research was supported by a grant of the Deutsche Forschungsgemeinschaft (DFG), awarded to Frank Goldhammer, Johannes Naumann, and Heiko Rölke (GO 1979/1-1). We are grateful to Beatrice Rammstedt and her group at GESIS (<http://gesis.org>), as well as the Federal Ministry for Education and Research (BMBF) for making the data available for this study.

Correspondence concerning this article should be addressed to Frank Goldhammer, German Institute for International Educational Research (DIPF), Schloßstr. 29, 60486 Frankfurt/Main, Germany. E-mail: [goldhammer@dipf.de](mailto:goldhammer@dipf.de)



teristics determines the way of cognitive processing. For instance, this can contribute to the validation of the assessment, if time on task can be related to the task response in a theoretically sound way.

Time on task is an important characteristic of the solution process indicating the duration of perceptual, cognitive, and psychomotorical activities. From a measurement point of view, the usefulness of time on task and the task outcome, respectively, depend on the tasks' difficulty. In easy tasks assessing basic skills, individual differences will mainly occur in response latencies, whereas accuracy will be consistently high. Following this logic, a number of assessment tools that address constructs like naming speed (e.g., Nicolson & Fawcett, 1994), visual word recognition (e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004), or number naming speed (e.g., Krajewski & Schneider, 2009) make use of time on task. In contrast, in more difficult tasks the accuracy of a result is of interest, for example, in assessments of reading comprehension (e.g., van den Broek, & Espin, 2012) or problem solving (e.g., Greiff, Wüstenberg, et al., 2013; Klieme, 2004; Mayer, 1994; Wirth & Klieme, 2003). In these skill assessments, time on task usually is not taken into account. Nevertheless, both the task result and time on task constitute task performance regardless of the task's difficulty.

In skill assessments, the relation between time on task and task result (accuracy) can be conceived of in two ways. On the one hand, taking more time to work on a task may be positively related to the result as the task is completed more thoroughly. On the other hand, the relation may be negative if working faster and more fluently reflects a higher skill level. The present study addresses these contradictory predictions and aims at clarifying the conditions of their validity by jointly analyzing task success and time on task data from the computer-based Programme for the International Assessment of Adult Competencies (PIAAC; cf. OECD, 2013; Schleicher, 2008). Thus, we take advantage of the fact that computer-based assessment renders data available on a large scale that was previously available only through small-scale experimenting (i.e., time on task). Data such as time spent on individual tasks can serve to answer basic research questions (such as clarifying the relation of time on task and task result in different domains). Furthermore, the data can enhance educational assessment. For instance, construct validation can be supported by testing whether behavioral process indicators are related to task outcomes as expected from theory.

### Time on Task

Time on task is understood as the time from task onset to task completion. Thus, if the task was completed in order, it reflects the time taken to become familiar with the task, to process the materials provided to solve the task, to think about the solution, and to give a response.<sup>1</sup> In tasks requiring the participant to interact with the stimulus through multiple steps, time on task can be further split into components, for instance, reflecting the time taken to process a single page from a multipage stimulus. To model time on task, two different approaches have been suggested (cf. van der Linden, 2007, 2009). First, time is considered an indicator of a (latent) construct, for example, reading speed (Carver, 1992) or reasoning speed (Goldhammer & Klein Entink, 2011). Here, response and time data are modeled using separate measurement

models. Second, within an explanatory item response model, time is used as a predictor to explain differences in task success (cf. Roskam, 1997). In the present study, this second approach is used to investigate the relation between time on task and task success. Task success (dependent variable) can be perceived as a function of time on task (independent variable) because the individual is able to control time spent on completing a task to some extent, which in turn may affect the probability of attaining the correct result (cf. van der Linden, 2009).

### Relation of Time on Task to Task Success

When investigating the relation between time on task and task success, the well-known speed–accuracy tradeoff, which is usually investigated in experimental research (cf. Luce, 1986), has to be taken into account. Tradeoff means that for a given person working on a particular task, accuracy will decrease as the person works faster. The positive relation between time on task and task success, as predicted by the speed–accuracy tradeoff, is a within-person phenomenon that can be expected for any task (e.g., Wickelgren, 1977). However, when switching from the within-person level to a population, the relation between time on task and task success might be completely different, for instance, a negative or no relation, although *within* each person, the speed–accuracy compromise remains as the positive relation between time on task and task success (cf. van der Linden, 2007). Consequently, at the population level, findings on the relation of time on task with task success may be heterogeneous. One line of research modeling time on task as an indicator of speed provides speed–skill or speed–ability correlations of different directions and strengths across domains. For example, for reasoning, positive correlations between skill (measured through task success) and slowness (measured through time on task) were found (e.g., Goldhammer & Klein Entink, 2011; Klein Entink, Fox, & van der Linden, 2009). For arithmetic zero correlations (van der Linden, Scrams, & Schnipke, 1999) were obtained, whereas for basic skills to operate a computer's graphical user interface, a negative relation was demonstrated (Goldhammer, Naumann, & Keßel, 2013), as was for basic reading tasks such as phonological comparison and lexical decision (Richter, Isberner, Naumann, & Kutzner, 2012).

These results suggest that the time on task effect might be moderated by domain and task difficulty. A comparison of tasks across studies reveals that in difficult tasks assessing for instance reasoning, task success is positively related to time on task, whereas in easy tasks, such as basic interactions with a computer interface, the relation is negative. Independent evidence for this line of reasoning comes from research suggesting that task difficulty within a given domain affects the association between time on task and task success. Neubauer (1990) investigated the correlation between the average time on task and the test score for figural reasoning tasks and found a zero correlation. However, for task clusters of low, medium, and high difficulty, he found nega-

<sup>1</sup> Depending on what is considered to be a task, there may be alternative definitions of *time on task*. For instance, in this special section, Kupiainen, Vainikainen, Marjanen, and Hautamäki (2014) use the term *time on task* to refer to the time needed to complete a test in a learning to learn assessment, whereas *response time* is considered to represent the time needed to respond to a single question or problem (which is comparable to our notion of time on task).



tive, zero, and positive correlations, respectively. Similarly, in a recent study by Dodonova and Dodonov (2013), the strength of the negative correlation between time on task and accuracy in a letter sequence task tended to decrease with increasing task difficulty.

### Time on Task Effects and Dual Processing Theory

An explanation for the heterogeneity of associations between time on task and task success may be provided by dual processing theory, which distinguishes between automatic and controlled mental processes (cf. Fitts & Posner, 1967; Schneider & Chein, 2003; Schneider & Shiffrin, 1977). Automatic processes are fast, proceduralized, and parallel; they require little effort and operate without active control or attention, whereas controlled processes are slow, are serial, require attentional control, and can be alternated quickly. Tasks are amenable to automatic processing due to learning only under consistent conditions, that is, rules for information processing including related information-processing components and their sequence are invariant (Ackerman, 1987). Learning under consistent conditions can be divided into three stages (cf. Ackerman & Cianciolo, 2000; Fitts & Posner, 1967). The first stage, when the individual acquires task knowledge and creates a production system (cf. Adaptive Control of Thought [ACT] theory; Anderson & Lebiere, 1998), is characterized by controlled processing. Automatic processing becomes more apparent in the second stage and dominates in the third stage. Thus, task performance is slow and error prone at the beginning of learning, but speed and accuracy increase as the strength of productions is increased through practice (Anderson, 1992).

Consequently, in domains and tasks that allow for automatic processing, a negative association between time on task and task success is expected. Well-practiced task completion is associated with both fast and correct responses. In contrast, a positive association is expected in domains and tasks that do not allow for a transition from controlled to automatic processing due to inconsistent processing rules and variable sequences of information processing. Taking more time to work carefully would positively impact task success. In line with this reasoning, Klein Entink et al. (2009) showed that test effort in a reasoning test, that is, the extent to which a test taker cares about the result, is positively related to test-taking slowness (measured through time on task), which itself is positively related to skill (measured through task success).

Notably, dual processing theory suggests a dynamic interaction of automatic and controlled processing in that the acquisition of higher level cognition is enabled by and builds upon automatic subsystems (Shiffrin & Schneider, 1977). Basically, tasks within and between domains are assumed to differ with respect to the composition of demands that necessarily require controlled processing and those that can pass into automatic processing (Schneider, & Fisk, 1983). Similarly, for a particular task, individuals are assumed to differ in the extent to which the task-specific information-processing elements that can be automatized are actually automatized (e.g., Carlson, Sullivan, & Schneider, 1989). In the following two sections, we describe in detail how automatic and controlled processes may interact in the two domains considered, reading and problem solving.

### Time on Task in Reading

Reading a text demands a number of cognitive component processes and related skills. Readers have to identify letters and words. Syntactic roles are then assigned to words, sentences are parsed for their syntax, and their meaning is extracted. Coherence must be established between sentences, and a representation of the propositional text base must be created, as well as a situation model of the text contents, integrated with prior knowledge (Kintsch, 1998). In addition, cognitive and metacognitive regulations might be employed. When text contents are learned, strategies of organization and elaboration will aid the learning process.

These different cognitive component skills allow for a transition from controlled to automatic processing to different degrees. Processes such as phonological recoding, orthographic comparison, or the retrieval of word meanings from long-term memory are slow and error prone in younger readers but become faster and more accurate as reading skill acquisition progresses (Richter, Isberner, Naumann, & Neeb, 2013). Indeed, theories of reading such as the lexical quality hypothesis (Perfetti, 2007) claim that reading skill rests on reliable as well as quickly retrievable lexical representations. In line with this, text comprehension is predicted by the speed of access to phonological, orthographic, and meaning representations (e.g., Richter et al., 2012, 2013). Beyond the word level, the speed of semantic integration and local coherence processes are equally positively related to comprehension (e.g., Naumann, Richter, Christmann, & Groeben, 2008; Naumann, Richter, Flender, Christmann, & Groeben, 2007; Richter et al., 2012). As shown by longitudinal studies, accuracy in reading assessments during primary school approaches perfection, whereas reading fluency reflecting reading performance per time unit continues to increase across years of schooling (cf. Landerl & Wimmer, 2008). The high accuracy rates suggest that reading is already well automatized during primary school.

Following this line of reasoning, in reading tasks, a negative time on task effect might be expected. A number of reading tasks, however, require attentional cognitive processing to a substantial degree as well. For instance, readers might need to actively choose which parts of a text to attend to when pursuing a given reading goal (e.g., Gräsel, Fischer, & Mandl, 2000; Naumann et al., 2007, 2008; Organisation for Economic Co-Operation and Development [OECD], 2011, chap. 3; Puntambekar & Stylianou, 2005). In the case of a difficult text, strategies such as rereading or engaging in self-explanations (e.g., Best, Rowe, Ozuru, & McNamara, 2005; McKeown, Beck, & Blake, 2009) are needed for comprehension. Also, in skilled readers, such processes require cognitive effort (Walczyk, 2000), and effort invested in strategic reading positively predicts comprehension (e.g., Richter, Naumann, Brunner, & Christmann, 2005; Sullivan, Gneddilow, & Puntambekar, 2011). This, however, will involve longer time spent on task.

Taken together, this means that in easy reading tasks, the potentially automatic nature of reading processes at the word, sentence, and local coherence level leads to a negative time on task effect (e.g., when reading a short and highly coherent linear text). As reading tasks become more difficult and readers need to engage in strategic and thus controlled cognitive processing, the negative time on task effect will be diminished or reversed.



## Time on Task in Problem Solving

Problem solving is required in situations where a person cannot attain a goal by using routine actions or thinking due to barriers or novelty (e.g., Funke & Frensch, 2007; Mayer, 1992; Wirth & Klieme, 2003). Problem solving requires higher order thinking, the finding of new solutions, and sometimes interaction with a dynamic environment (Klieme, 2004; Mayer, 1994). In the present study, a specific concept of problem solving as defined for the PIAAC study is taken into account; it refers to solving information problems in technology-rich environments. That is, technology-based tools and information sources (e.g., search engines, Web pages) are used to solve a given problem by “storing, processing, representing, and communicating symbolic information” (OECD, 2009b, p. 8). Information problems in this sense (e.g., finding information on the Web fulfilling multiple criteria to take a decision) cannot be solved immediately and routinely. They require developing a plan consisting of a set of properly arranged subgoals and performing corresponding actions through which the goal state can be reached (e.g., identifying the need for information to be obtained from the Web, defining an appropriate Web search query, scanning the search engine results page, checking linked Web pages for multiple criteria, collecting and comparing information from selected Web pages, and making use of it in the decision to be taken). This differs, for instance, from solving logical or mathematical problems where complexity is determined by reasoning requirements but not primarily by the information that needs to be accessed and used (OECD, 2009b). Cognitive and metacognitive aspects of problem solving as assessed in PIAAC include setting up appropriate goals and plans to achieve the goal state. This includes monitoring the progress of goal attainment, accessing and evaluating multiple sources of information, and making use of this information (OECD, 2009b, p. 11).

Problem solving is a prototype of an activity that relies on controlled processing. Controlled processing enables an individual to deal with novel situations for which automatic procedures and productions have not yet been learned. Otherwise, the situation would not constitute a problem. Accordingly, Schneider and Fisk (1983) described skilled behavior in problem solving and strategy planning as a function of controlled processing. Notably, problem solving skill may also benefit from practice. The development of fluent component skills at the level of subgoals enables problem solvers to improve their strategies optimizing the problem solving process (see, e.g., Carlson, Khoo, Yaure, & Schneider, 1990).

General conceptualizations of (complex) problem solving conceive problem solving performance as consisting of knowledge acquisition including problem representation and the application of this knowledge to generate solutions (cf. Funke, 2001; Greiff, Wüstenberg, et al., 2013). Wirth and Leutner (2008) identified two simultaneous goals in the knowledge acquisition phase, that is, generating information through inductive search and integrating this information into a coherent model. Successful problem solvers move more quickly from identification to integration and thus will be able to invest time in advanced modeling and prediction (which provide the basis for successful knowledge application) rather than in low-level information processing.

Problem solving in technology-rich environments assumes two concepts, accessing information and making use of it, that seem similar to knowledge acquisition and application. However, there

are differences in that, for instance, retrieving information (e.g., by means of a search engine) is not comparable to an inductive search for rules governing an unknown complex system. Nevertheless, the various notions of problem solving assume successive steps of controlled information processing that may benefit from fluent component skills.

Therefore, a positive effect of time on task on task success is expected for problem solving. Taking sufficient time allows for all serial steps to planned subgoals to be processed, as well as more sophisticated operations to be used and properly monitored regarding progress. Particularly for weak problems solvers, spending more time on a task may be helpful to compensate for a lack of automaticity in required subsystems (e.g., reading or computer handling processes).

## Research Goal and Hypotheses

Our general research goal was to assess and investigate behavioral processes and their relation to task performance in computer-based assessment. More specifically, we determined the effect of time on task on the task result and the conditions that influence the strength and direction of this effect. For this, we used the computer-based assessment of reading and problem solving in the international large-scale study PIAAC, including log file data generated by the assessment system.

From a dual processing framework, we derived the general hypothesis that the relative degree of controlled versus automatic cognitive processing as required by a task, as well as the test taker's acquired skill level, determines the strength and direction of the time on task effect. The following three hypotheses address time on task effects across domains, task properties, and person characteristics. The fourth hypothesis aims at validating the interpretation of the time on task effect in problem solving by splitting up the global time on task into components that represent different steps of task solution and information processing.

*Hypothesis 1: Time on task effect across domains.* We expected a positive time on task effect for problem solving in technology-rich environment tasks. A negative time on task effect was expected for reading tasks because, in reading tasks, a number of component cognitive processes are apt for automatization. Problem solving, in contrast, by definition must rely on controlled processing to a substantial degree in each task.

*Hypothesis 2: Time on task effect across tasks.* Within domains, we expected the time on task effect to be moderated by task difficulty. Easy tasks can be assumed to be completed substantially by means of automatic processing, whereas difficult tasks evoking more errors require a higher level of controlled processing. Accordingly, we expected a positive time on task effect in problem solving to be accelerated with increasing task difficulty, and a negative time on task effect in reading to diminish with increasing task difficulty.

As our interpretation of the time on task effect focuses the way of cognitive processing, we additionally explored the potentially moderating role of the cognitive operation involved in each task as defined a priori by the PIAAC assessment framework (e.g., access in reading). More specifically, we investigated whether the task characteristic “cognitive



operation” explains task difficulty and if so whether the time on task effect would depend on the presence of specific cognitive operations.

*Hypothesis 3: Time on task effect across persons.* For a given task, individuals are assumed to differ in the extent to which the information-processing elements that are amenable to automatic processing are actually automatized. Highly skilled individuals are expected to be in command of well-automatized procedures within task solution subsystems that are apt to automatization (such as decoding in reading or using shortcuts to perform basic operations in a computer environment). We therefore expect the time on task effect to vary across persons. On the one hand, we predict that the time on task effect gets more positive for less skilled problem solvers and less negative for less skilled readers since they are expected to accomplish tasks with higher demands of controlled and strategic processing than skilled persons. For example, poor readers may rely on compensatory behaviors and strategies, especially when completing difficult tasks (see Walczyk, 2000). On the other hand, for skilled persons, we expect the inverse result, that is, due to a higher degree of routinized processing, the time on task effect gets less positive for skilled problem solvers and more negative for skilled readers.

*Hypothesis 4: Decomposing time on task effect at task level.* Computer-based assessment and especially the exploitation of log file data can help to further understand the task completion process. By moving from the global process measure of time on task to the underlying constituents, we can further validate the interpretation of the time on task effect. This is especially true for tasks requiring a complex sequence of stimulus interactions that can be reconstructed from a log file, giving insight into the accuracy and timing with which subgoals were being completed. In the present study, tasks assessing problem solving in technology-rich environments are highly interactive, requiring the operation of simulated computer and software environments or navigation in simulated Web environments. For a particular task, we expect that a positive time on task effect is confined to the completion of steps that are crucial for a correct solution (e.g., in a Web environment, visiting a page that presents information needed to give a correct response), whereas for others the effect is assumed to be negative (e.g., in a Web environment, visiting an irrelevant page). If this were the case, it would corroborate our assumption that it is the need for strategic and controlled allocation of cognitive resources that produces a positive time on task effect in problem solving or very difficult reading tasks.

## Method

### Sample

The PIAAC study initiated internationally by the OECD (cf. OECD, 2013; Schleicher, 2008) is a fully computer-based international comparative study assessing the competence levels of adults in 2011–2012. For the present study, data provided by GESIS–Leibniz Institute for the Social Sciences from the German

PIAAC field test in 2010 were used. The target population consisted of all noninstitutionalized adults between the ages of 16 and 65 years (inclusive) who resided in Germany at the time of sample selection and were enrolled in the population register. For the field test in Germany, a three-stage sampling was used with probability sample of communities and individuals in five selected federal states. The within-household sample included in the present study comprised 1,020 individuals completing the computer-based PIAAC assessment. Of these, 520 were male (50.98%) and 458 female (44.90%). For 42 participants, no gender information was available (4.12%). The average age was 39.40 years ( $SD = 13.30$ ).

### Instrumentation

**Reading literacy.** The PIAAC conceptual framework for reading literacy is based on conceptions of literacy from the International Adult Literacy Survey (IALS) conducted in the 1990s and the Adult Literacy and Life Skills Survey (ALL) conducted in 2003 and 2006 (see OECD, 2009a). It was extended for PIAAC to cover reading skill in the information age by including skills of reading in digital environments. More than half of the reading tasks were taken from the former paper-based adult literacy assessments IALS and ALL to link PIAAC results back to these studies. New tasks simulating digital (hypertext) environments were developed to cover the broadened construct including skills of reading digital texts. The tasks covered the cognitive operations “access and identify information,” “integrate and interpret information,” and “evaluate and reflect information” (see OECD, 2009a). The majority of tasks included print-based texts as used in previous studies (e.g., newspapers, magazines, books). Tasks representing the digital medium included, for instance, hypertext and environments such as message boards and chat rooms. Tasks are also varied with respect to the context (e.g., work/occupation, education and training) and whether they included continuous texts (e.g., magazine articles), noncontinuous texts (e.g., tables, graphs), or both.

In the PIAAC field test, 72 reading tasks were administered. For the present study, only those 49 tasks were used that entered the main study. To respond, participants were required to highlight text, to click a (graphical) element of the stimulus, to click a link, or to select a check box. As a sample, Figure 1 (upper panel) presents a screenshot from the first “Preschool Rules” task. Respondents were asked to answer the question shown on the left side of the screen by highlighting text in the list of preschool rules on the right side. The question was to figure out the latest time that children should arrive at preschool. Thus, readers were required to access and identify information, the context was personal, and print text was presented.

**Problem solving in technology-rich environments.** This construct refers to using information and communication technology (ICT) to collect and evaluate information so as to communicate and perform practical tasks such as organizing a social activity, deciding between alternative offers, or judging the risks of medical treatments (OECD, 2009b). The framework (OECD, 2009b) defined multiple task characteristics that formed the basis for instrument development. The cognitive operations to be covered by the tasks were goal setting and progress monitoring, planning and self-organizing, acquiring and evaluating information, and making use of information. The technology dimensions



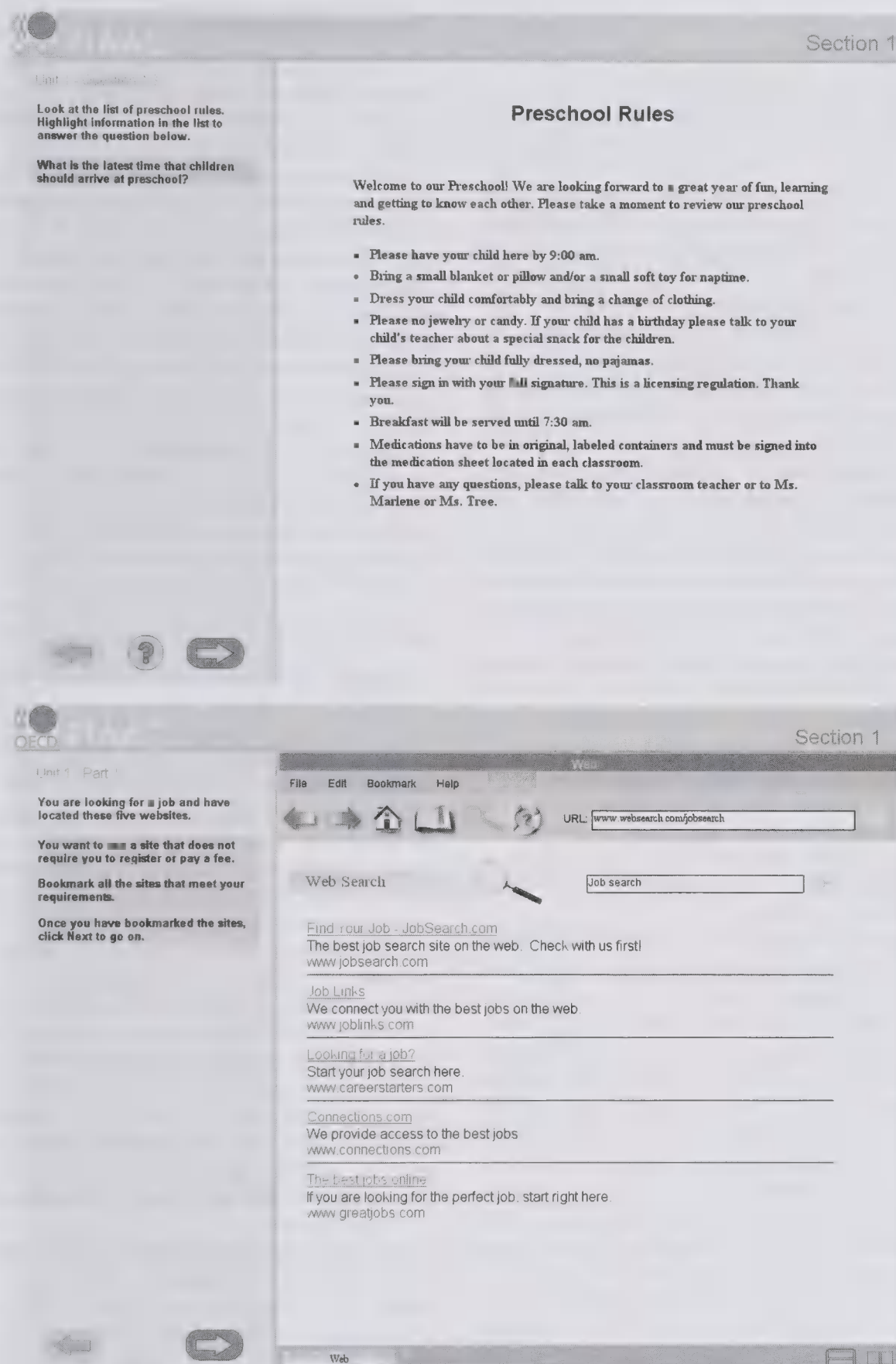


Figure 1. Sample tasks: reading literacy task "Preschool Rules" (upper panel); problem solving in technology-rich environments task "Job Search" with only the start page showing the search engine results depicted, not the linked pages (lower panel). OECD = Organisation for Economic Co-Operation and Development; PIAAC = Programme for the International Assessment of Adult Competencies.

included hardware devices (e.g., desktop or laptop computers), software applications (e.g., file management, Web browser, e-mail, spreadsheet), various commands and functions (e.g., but-

tons, links, sort, find), and multiple representations (e.g., text, numbers, graphics). Moreover, task development aimed at the variation of the task's purpose (e.g., personal, work/occupation),



intrinsic complexity (e.g., the minimal number of actions required to solve the problem, the number of constraints to be satisfied), and the explicitness of the problem (implicit, explicit).

As defined by the framework (OECD, 2009b), tasks were developed in such a way that they varied in the number of required cognitive operations (e.g., acquiring and evaluating information), the number and kind of actions that have to be taken to solve the task in a computer environment, the inclusion of unexpected outcomes or impasses, and the extent to which the tasks were open-ended. A more difficult task simulating real-life problem solving would require several cognitive operations, multiple actions in different environments, unexpected outcomes, and the planning of multiple subgoals that may depend on each other. A corresponding sample task would be one in which the problem solver has to do a Web search on the Internet to access information, integrate and evaluate information from multiple online sources by using a spreadsheet, and then create a summary of the information to be presented at school by using a presentation software.

In the PIAAC field test, 24 problem solving tasks were administered. Of these tasks, only 13 were selected for the main study. For the present study, all available tasks were considered to obtain more reliable results on the correlation of effects varying across tasks. After excluding tasks with poor discrimination and tasks for which no score could be derived, 18 tasks were left. In the context of international large-scale assessments, further tasks may be dropped, especially if they show differential item functioning across participating countries. However, as we only used national data and did not aim at comparing countries, there was no need to consider task-by-country interactions. To give a response in the simulated computer environments, participants were required to click buttons, menu items, or links, to select from drop-down menus, to drag and drop, and so on.

As a sample, Figure 1 (lower panel) presents a screenshot from the task "Job Search." Regarding cognitive operations, participants had to access and evaluate information and monitor criteria for constraint satisfaction within a simulated job search. Thus, the task's purpose was occupational. Starting from a search engine results page, the task was to find all the sites that do not require users to register or pay a fee and to bookmark these sites. Regarding the explicitness of the problem, instructions did not directly tell participants the number of sites they must locate, but evaluation criteria were clearly stated. To solve the task, single actions of evaluation had to be repeated for each website; for a target page, multiple constraints needed to be satisfied. Both characteristics determined intrinsic complexity. As regards software applications and related commands, the task was situated in a simulated Web environment that included tools and functionality similar to those found in real-life browser applications, that is, clickable links, back and forward buttons of the browser, and a bookmark manager that allowed one to create, view, and change bookmarks. The opening page presented the task description on the left side and the results of the Web search engine, that is, clickable links and brief information about the linked page, on the right side of the screen. From this search engine results page, participants had to access the hypertext documents connected via hyperlinks to locate and bookmark those websites that meet the search criteria.

## Design and Procedure

A rotation design was used to form 21 booklets resulting in an effective sample size for reading literacy of 113 to 146 responses per task and for problem solving in technology-rich environments of 140 to 191 responses per task.

Data were collected in computer-assisted personal interviews. Interviewers went to the participants' households to conduct the interview in person. First, participants completed a background questionnaire, and then the interviewer handed the notebook to the participant for completion of the cognitive tasks. There was no global time limit, that is, participants could take as long as they needed. Participants only completed the computer-based tasks if they were sufficiently ICT literate, which was tested by ICT tasks requiring basic operations such as highlighting text by clicking and dragging. In case of nonsufficient ICT literacy, a paper-based assessment was administered. In the computer-based part, participants were randomly assigned to booklets including reading literacy, numeracy, and problem solving tasks. For the present study, only data from the computer-based assessment of reading literacy and problem solving were included.

## Statistical Analyses

**Modeling approach.** The generalized linear mixed model (GLMM) framework (e.g., Baayen, Davidson, & Bates, 2008; De Boeck et al., 2011; Doran, Bates, Bliese, & Dowling, 2007) was used to investigate the role of time on task in reading and problem solving (Hypotheses 1–3). A linear model consists of a component  $\eta_{pi}$ , representing a linear combination of predictors determining the probability of person  $p$  for solving task  $i$  correctly. The predictors' weights are called effects. Modeling mixed effects means to include both random effects and fixed effects. Fixed effects are constants across units or groups of a population (e.g., tasks, persons, classrooms), whereas random effects may vary across units or groups of a population (cf. Gelman, 2005). The generalized version of the linear mixed model accommodates also categorical response variables. In measurement models of item response theory, for instance, the effect of each item or task  $i$  on the probability of obtaining a correct response is typically estimated as a fixed effect representing the task's difficulty or easiness. The effect of person  $p$  is usually modeled as random, that is, as an effect which may vary across persons and for which the variance is estimated. The variance of this random effect represents the variability of skill across persons.

The GLMM incorporating both random effects,  $\mathbf{b}$ , and fixed effects,  $\boldsymbol{\beta}$ , can be formulated as follows:  $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$  (e.g., Doran et al., 2007). In this model,  $\mathbf{X}$  is a model matrix for predictors with fixed weights included in vector  $\boldsymbol{\beta}$ , and  $\mathbf{Z}$  is a model matrix for predictors with random weights included in vector  $\mathbf{b}$ . The distribution of the random effects is modeled as a multivariate normal distribution,  $\mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ , with  $\boldsymbol{\Sigma}$  as the covariance matrix of the random effects. The continuous linear component  $\eta_{pi}$  is linked to the observed ordered categorical response  $Y_{pi}$  (correct vs. incorrect) by transforming the expected value of the observed response, that is, the probability to obtain a correct response  $\pi_{pi}$ . When using the log-transformed odds ratio (log-odds), the logit link function follows:  $\eta_{pi} = \ln(\pi_{pi}/(1 - \pi_{pi}))$  (cf. De Boeck et al., 2011).



In the present study, to address the research question of whether the strength of the time on task effect is correlated with the easiness of tasks, the effects of both persons and tasks were defined as random intercepts (cf. random person random item model; De Boeck, 2008). A fixed intercept,  $\beta_0$ , is estimated additionally, which is the same for all participants and tasks.

A baseline Model M0 was obtained by specifying an item response model (1PL or Rasch model) with task and person as random intercepts and by adding the time on task as person-by-item predictor with a fixed effect  $\beta_1$ . Model M0 serves as parsimonious reference model that is compared with more complex models including further fixed and/or random effects:  $\eta_{pi} = (\text{intercept } \beta_0) + (\text{individual skill } b_{op}) + (\text{relative easiness } b_{oi}) + \beta_1$  (time on task  $t_{pi}$ ).

In the following analyses, this model is systematically extended by adding further predictors. For example, the predictor (time on task  $t_{pi}$ ) with the random weight  $b_{1i}$  is added, providing the variance of the by-task adjustment  $b_{1i}$  to the fixed time on task effect  $\beta_1$ . As the by-task adjustment,  $b_{1i}$ , and task easiness,  $b_{oi}$ , are tied to the same observational unit, that is, task  $i$ , their association is also estimated. This correlation can be used to test whether the strength of the time on task effect linearly depends on task difficulty (as claimed by Hypothesis 2). Figure 2 shows the path diagram of Model M1, which is Model M0 extended by the predictor (time on task  $t_{pi}$ ) with a random weight across tasks,  $b_{1i}$  (cf. the graphical representations of GLMMs by De Boeck & Wilson, 2004). In Model M1, there is a fixed time on task effect,  $\beta_1$ , representing the average time on task effect. However, it is

adjusted by task by adding the weight  $b_{1i}$ , which allows the time on task effect to vary across tasks as indicated by subscript  $i$ . The other models under consideration can be derived in a similar fashion by adding random effects adjusting the time on task effect by cognitive operation (Model M2, cf. Hypothesis 2), by person (Model M3, cf. Hypothesis 3), or by task and person (Model M4, integrating Hypothesis 2 and Hypothesis 3).

To clarify whether the introduction of further random components into the model significantly improves model fit, model comparison tests were conducted. For comparing nested models, the likelihood ratio (LR) test was used, which is appropriate for inference on random effects (Bolker et al., 2009). The test statistic, that is, twice the difference in the log-likelihoods, is approximately  $\chi^2$  distributed with degrees of freedom equal to the number of extra parameters in the more complex model. The LR test is problematic when the null hypothesis implies the variance of a random effect to be zero; this means that the parameter value is on the boundary of the parameter space (boundary effect; cf. Baayen et al., 2008; Bolker et al., 2009; De Boeck et al., 2011). Using the chi-square reference distribution increases the risk of Type II errors; therefore, the LR test has to be considered as a conservative test for variance parameters.

For the analysis at the task level (Hypothesis 4), logistic regression was used to predict task success by the time taken on individual steps of the task completion sequence.

**Interpreting the effect of time on task in the GLMM.** The “fundamental equation of RT modeling” (van der Linden, 2009, p. 259) assumes that the response time (RT; time on task) of person

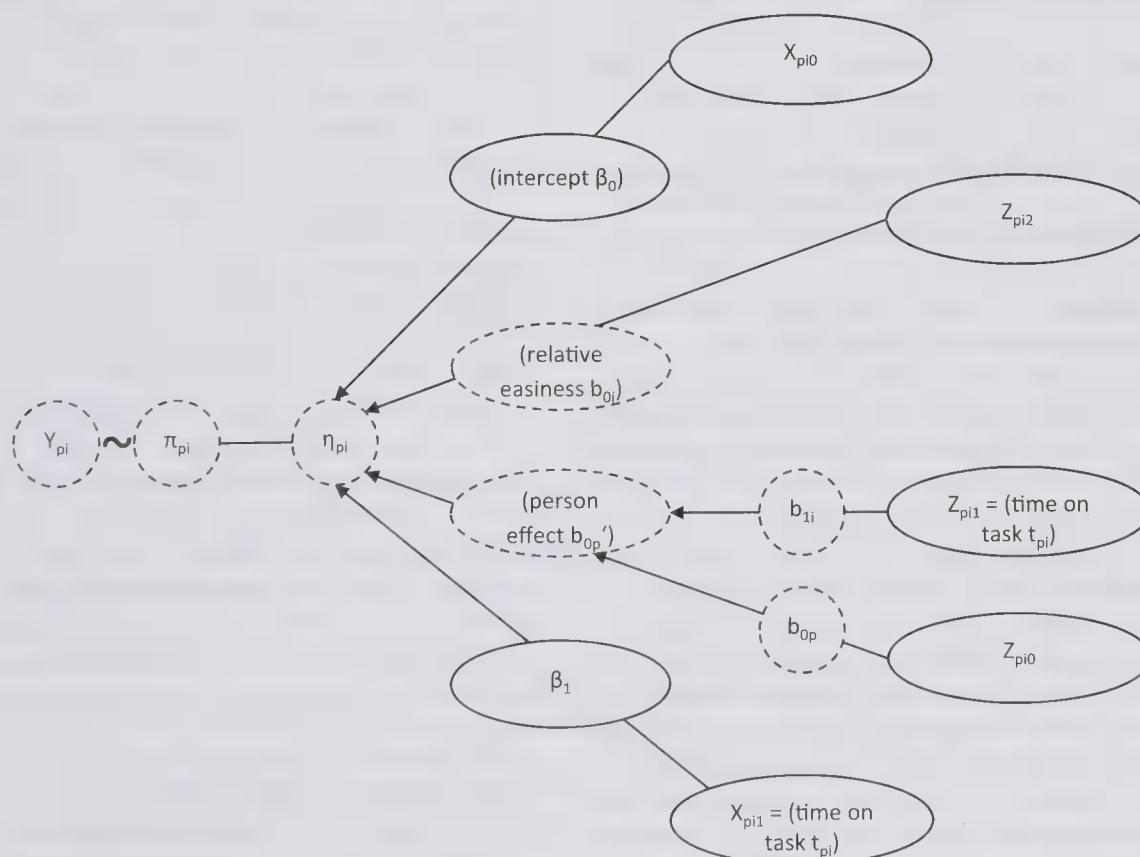


Figure 2. Graphical representation of model M1 showing how the probability to obtain a correct response,  $\eta_{pi}$ , is affected by a general intercept,  $\beta_0$ , the relative task easiness,  $b_{oi}$ , and individual skill,  $b_{op}$ . Moreover, there is a time on task effect consisting of a fixed part,  $\beta_1$ , as well as random part,  $b_{1i}$ , which means that the time on task effect may vary across tasks  $i$ .



$p$  when completing task  $i$  depends both on the person's speed  $\tau_p$  and the task's time intensity  $\lambda_i$ . Accordingly, the expected value of the (log-transformed) response time can be defined as follows:  $E(\ln(t_{pi})) = \lambda_i - \tau_p$  (cf. van der Linden, 2009). This implies that the effect of time on task reflects both the effect of the person and the task component.

When the effect of time on task is introduced as an overall fixed effect  $\beta_1$ , as in Model M0, this effect would reflect the association between time on task and the log-odds ratio of the expected response. This association could not be interpreted in a straightforward way, as it depends not only on the correlation between underlying person-level parameters, that is, skill and speed, but also on the correlation of corresponding item parameters, that is, difficulty and time intensity (see van der Linden, 2009).<sup>2</sup> However, when modeling the effect of time on task as an effect random across tasks (Hypothesis 1), groups of tasks supposed to be homogeneous (Hypothesis 2), or individuals (Hypothesis 3), the influences from the task and person levels can be disentangled.

A time on task effect random across tasks is obtained by introducing the by-task adjustment  $b_{1i}$  to the fixed time on task effect  $\beta_1$ . The time on task effect by task results as  $\beta_1 + b_{1i}$ . Thereby, time on task is turned into a person-level covariate varying between tasks. That is, given a particular task with certain time intensity, variation in time on task is only due to differences in persons' speed (plus residual). This allows us to interpret time on task as an task-specific speed parameter predicting task success above and beyond individual skill.

A by-person random time on task effect means to adjust the fixed time on task effect  $\beta_1$  by the person-specific parameter  $b_{1p}$ , resulting in the time on task effect  $\beta_1 + b_{1p}$ . The fixed effect shows a constant as subscript, whereas the random effect is provided additionally with  $p$  as subscript indicating that the effect may vary across persons  $p$ . Given a particular person working at a certain speed level, variation in time on task is only due to differences in the tasks' time intensity (plus residual). This means that time on task can be conceived of as a task-level covariate that is specific to persons and predicts task success above and beyond task easiness.

**Trimming of time data.** As a preparatory step for data analysis, the (between-person) time on task distribution of each task was inspected for outliers. The middle part of a time on task distribution was assumed to include the observations that are most likely to come from the cognitive processes of interest. To exclude extreme outliers in time on task and to minimize their effect on analyses, observations two standard deviations above (below) the mean were replaced by the value at two standard deviations above (below) the mean. As even a single extreme outlier can considerably affect mean and standard deviation, time on task values were initially log-transformed, which means that extremely long time on task values were pulled to the middle of the distribution. With this trimming approach, 4.79% of the data points in reading literacy and 4.67% in problem solving were replaced. Transforming a covariate may have an impact on estimated parameters of the linear mixed model (for linear transformations, see, e.g., Morrell, Pearson, & Brant, 1997). Therefore, we conducted the analyses also without log-transforming the time on task variable. As we obtained the same result pattern, we report the analyses with log transformation only. Results obtained with the untransformed data are available from the first author upon request.

**Statistical software.** For estimating the presented GLMMs, the lmer function of the R package lme4 (Bates, Maechler, & Bolker, 2012) was applied. The R environment (R Core Team, 2012) was also used to conduct logistic regression analyses.

## Results

### Difficulty of Tasks

To compare the difficulty of problem solving tasks and reading literacy tasks, the baseline Model M0 was tested for both domains without the time on task effect. For reading literacy, an intercept of  $\beta_0 = 0.61$  ( $z = 3.21$ ,  $p < .01$ ) was obtained; it represents the marginal log-odds for a correct response in a task of average easiness completed by a person of average skill; the corresponding probability was 64.68%. For problem solving, the result was  $\beta_0 = -0.72$  ( $z = -2.37$ ,  $p < .01$ ), indicating that the probability of a correct response was on average only 32.68%, that is, problem solving tasks were much harder than reading literacy tasks. Figure 3 shows the densities of the estimated task easiness parameters for reading literacy tasks (upper panel) and problem solving tasks (lower panel). Task easiness values were obtained by adding the intercept  $\beta_0$  and the random task intercept (relative easiness  $b_{0i}$ ). The proportion of correct responses,  $p$ , ranged for reading literacy from 12.41% to 96.92% and for problem solving from 11.86% to 77.49%.

### Time on Task Effect by Domain (Hypothesis 1)

For testing Hypotheses 1 and 2, Model M0 was extended to Model M1 by adding the by-task random time on task effect  $b_{1i}$ :  $\eta_{pi} = (\text{intercept } \beta_0) + (\text{individual skill } b_{0p}) + (\text{relative easiness } b_{0i}) + \beta_1 (\text{time on task } t_{pi}) + b_{1i} (\text{time on task } t_{pi})$ .

To address Hypothesis 1 regarding the time on task effect by domain, the fixed time on task effects  $\beta_1$ , as specified in Model M1 (see also Figure 2), were compared between reading literacy and problem solving.

**Reading literacy.** Table 1 provides an overview of the results. For reading literacy, a negative and significant time on task effect of  $\beta_1 = -0.61$  ( $z = -4.90$ ,  $p < .001$ ) was found. Thus, for a reading literacy task of average difficulty, correct responses were associated with shorter times on task, whereas incorrect responses were associated with longer times on task.

**Problem solving.** For problem solving, a positive and significant time on task effect of  $\beta_1 = 0.56$  ( $z = 2.30$ ,  $p = .02$ ) was estimated. Thus, for a problem solving task of average difficulty, correct responses were associated with longer times on task and vice versa. These findings give support to Hypothesis 1.

### Time on Task Effect by Task (Hypothesis 2)

If the assumption holds that task difficulty moderates the time on task effect, a relation between task easiness and the strength of the time on task effect should be observable within a domain. To test Hypothesis 2, the variances of the by-task adjustments to the fixed time on task effects and their correlations with task easiness,

<sup>2</sup> We thank an anonymous reviewer who advised us to consider this issue.



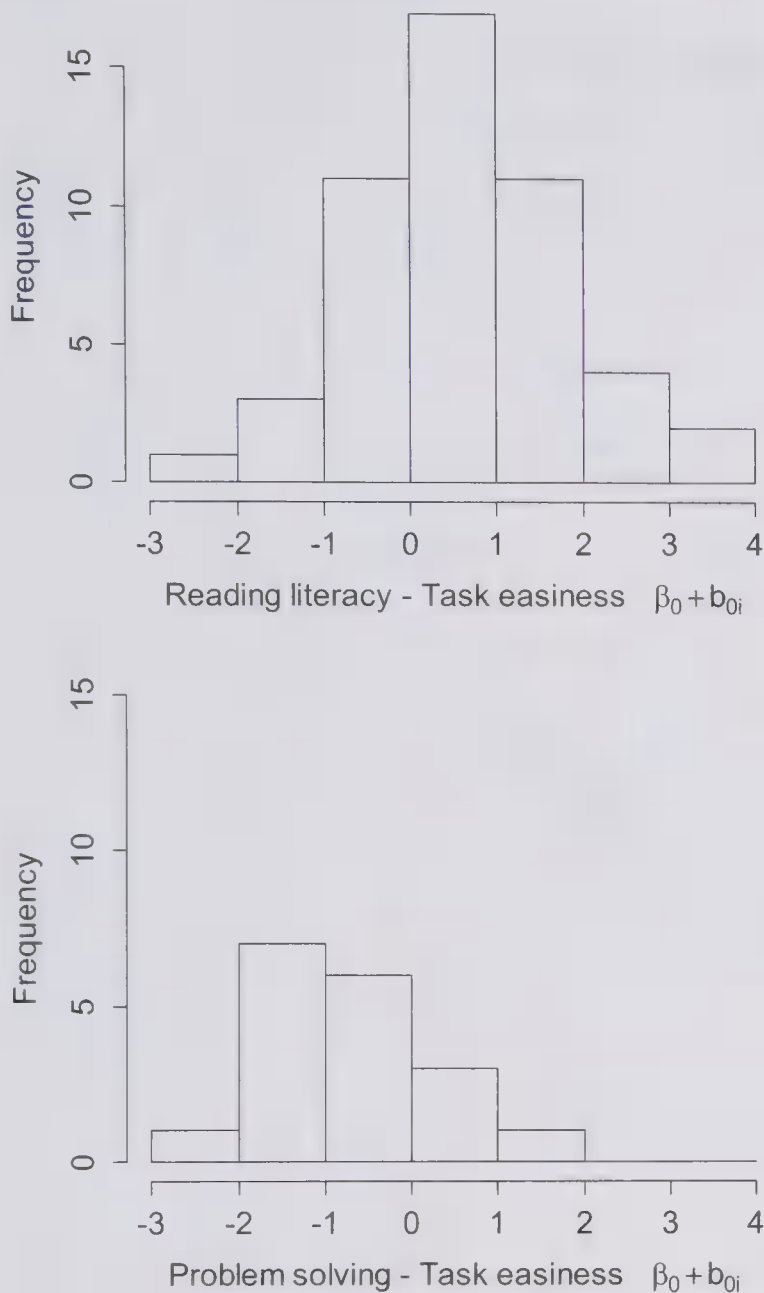


Figure 3. Distribution of estimated task easiness parameters for reading literacy (upper panel) and problem solving in technology-rich environments (lower panel). On average, reading literacy tasks were easier than problem solving tasks.

as estimated through Model M1, were inspected for both domains under consideration.

**Reading literacy.** For reading literacy, the variability of the by-task adjustment was estimated to be  $\text{Var}(b_{1i}) = 0.55$ . This means that for reading literacy, the time on task effect varied across tasks. Most importantly, the by-task time on task effect and intercept were negatively correlated,  $\text{Cor}(b_{0i}, b_{1i}) = -.39$ . That is, the overall negative time on task effect became even stronger in easy tasks but was attenuated in difficult tasks. The upper left panel in Figure 4 illustrates how the time on task effect in reading literacy was adjusted by task. To test whether the model extension improved the model's goodness of fit, we compared the nested Models M0 and M1. The difference test showed that Model M1 fitted the data significantly better than Model M0,  $\chi^2(2) = 77.65$ ,  $p < .001$ . To test whether the correlation parameter was actually needed to improve model fit, that is, to test the significance of the

correlation, Model M1 was compared to a restricted version (Model M1r), which did not assume a correlation between by-task time on task effect and by-task intercept. The model difference test suggested that the unrestricted version of Model M1 had a better fit to the data than the restricted version,  $\chi^2(1) = 5.16$ ,  $p = .02$ . Thus, the negative correlation between the by-task adjustment of the time on task effect and the random task intercept (i.e., task easiness) was also significant.

**Problem solving.** For problem solving, the variance of the by-task adjustment to the fixed effect of time on task was estimated as  $\text{Var}(b_{1i}) = 0.89$ . Thus, for problem solving in technology-rich environments, the time on task effect varied across tasks. The correlation between the by-task adjustment to the time on task effect and task easiness was negative as for reading literacy,  $\text{Cor}(b_{0i}, b_{1i}) = -.61$ . That is, the overall positive time on task effect became even stronger in hard-to-solve tasks but was attenuated in easy-to-solve tasks. Figure 4 (upper right panel) illustrates how the time on task effect in problem solving was adjusted by task. The model difference test, comparing the nested Models M0 and M1, clearly showed that adding the random time on task effect in Model M1 improved the model fit,  $\chi^2(2) = 73.99$ ,  $p < .001$ . Moreover, comparing Model M1 with a restricted version (Model M1r) without a correlation between the by-task time on task effect and the random task intercept revealed that the correlation was significant,  $\chi^2(1) = 6.50$ ,  $p = .01$ .

All together, these results give clear support to Hypothesis 2. In a domain where task solution cannot rely on automatic processes such as problem solving, the already positive time on task effect was substantially increased in tasks that were especially difficult. In a domain where rapid automatic processing can account for a substantial part of the task solution process such as reading, an already negative time on task effect became even stronger in easier tasks but diminished in more difficult tasks.

### Time on Task Effect by Cognitive Operation

An alternative explanation for the variability of the time on task effect between tasks refers to differences in the required cognitive operations. That is, tasks being homogeneous with respect to cognitive operations would show similar time on task effects. To test whether the presence of different cognitive operations as detailed by the respective frameworks affects the time on task effect, we extended Model M0 to the following Model M2 by introducing the cognitive operation  $c$  required in a task as a categorical task-level predictor and as a factor moderating the time on task effect, which is represented by the random weight  $b_{1c}$ :  $\eta_{pi} = (\text{intercept } \beta_0) + (\text{individual skill } b_{0p}) + (\text{relative easiness } b_{0i}) + \beta_1 (\text{time on task } t_{pi}) + (\text{cognitive operation } b_{0c}) + b_{1c} (\text{time on task } t_{pi})$ .

**Reading literacy.** For reading literacy, the PIAAC framework assumes three broad aspects of cognitive operation, access and identify information, integrate and interpret information, and evaluate and reflect information. In a first step, we tested an explanatory item response model with random person and task effects as well as the effect of cognitive operation. For the three aspects of cognitive operations, the intercepts of 1.07 ( $z = 4.72$ ,  $p < .01$ ), 0.00 ( $z = 0.00$ ,  $p = 1.00$ ), and 0.08 ( $z = 0.19$ ,  $p = .85$ ) were estimated. The probabilities of a correct response corresponding to these intercepts were 74.50%, 50.01%, and 51.96%. Access tasks



Table 1

Overview of Main Model Parameters on the Time on Task Effect and Model Comparison Tests

Domain	Research question/hypothesis	Model	Time on task effect random across	$\chi^2$ of model difference test ( <i>df</i> in parentheses)	Fixed-effect $\beta_1$	Variance of random effect	Correlation of random effects
Reading literacy	Baseline model	M0			−0.55***	—	—
	Testing Hypotheses 1 and 2: Time on task effect by domain and task	M1	Tasks		−0.61***	0.55	−.39
	Comparison with baseline model	M1 vs. M0		77.65 (2)***			
	Restricted model without random effect correlation	M1r	Tasks		−0.59***	0.54	—
	Comparison with unrestricted model	M1 vs. M1r		5.16 (1)*			
	Exploring the time on task effect by cognitive operation	M2	Cognitive operations		−0.51***	0.003	−1.00
	Restricted model without random time on task effect across cognitive operations	M2r			−0.55***	—	—
	Comparison with unrestricted model	M2 vs. M2r		0.79 (1), <i>ns</i>			
	Testing Hypothesis 3: Time on task effect by person	M3	Persons		−0.65***	0.14	−.65
	Comparison with baseline model	M3 vs. M0		15.09 (2)**			
	Restricted model without random effect correlation	M3r	Persons		−0.57***	0.09	—
	Comparison with unrestricted model	M3 vs. M3r		12.85 (1)**			
	Integrated model: Time on task effect by task and person	M4	Tasks Persons		−0.69**	0.64 0.23	−.52 −.78
	Comparison with baseline model	M4 vs. M0		106.14 (4)***			
Problem solving	Baseline model	M0			0.49***	—	—
	Testing Hypotheses 1 and 2: Time on task effect by domain and task	M1	Tasks		0.56*	0.89	−.61
	Comparison with baseline model	M1 vs. M0		73.99 (2)***			
	Restricted model without random effect correlation	M1r	Tasks		0.54*	0.87	—
	Comparison with unrestricted model	M1 vs. M1r		6.50 (1)*			
	Testing Hypothesis 3: Time on task effect by person	M3	Persons		0.51***	0.22	−.79
	Comparison with baseline model	M3 vs. M0		5.98 (2) <sup>†</sup>			
	Restricted model without random effect correlation	M3r	Persons		0.49***	0.12	—
	Comparison with unrestricted model	M3 vs. M3r		5.98 (1)*			
	Integrated model: Time on task effect by task and person	M4	Tasks Persons		0.56*	0.89 0.11	−.63 −.76
	Comparison with baseline model	M4 vs. M0		76.77 (4)***			

Note. Dashes indicate that a parameter was not included in the model. M = Model; r = restricted. *ns* = not significant.

<sup>†</sup>  $p < .10$ . \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

were thus relatively easy, whereas integrate and evaluate tasks show quite the same level of medium difficulty; by introducing cognitive operation as an explanatory variable of task easiness, the variance of task easiness,  $\text{Var}(b_{0i})$ , decreased from 1.52 to 1.24, which corresponds to  $R^2 = .20$ .

To investigate whether the influence of time on task on task success varies across cognitive operations, Model M2 was tested. The obtained variance of the by-cognitive operation adjustment to the time on task effect was only  $\text{Var}(b_{1c}) = 0.003$ . Moreover, the correlation with the corresponding intercept was  $\text{Cor}(b_{0c}, b_{1c}) = -1.00$ , indicating overparameterization of the model. Model M2

was compared with a restricted model including no time effect varying across cognitive operations (Model M2r); there was no significant improvement of model fit,  $\chi^2(2) = 0.79$ ,  $p = .67$ . Thus, the time on task effect did not vary across cognitive operations.

**Problem solving.** The time on task effect was not further investigated with respect to cognitive operations for two reasons. First, there was only a small set of 18 tasks available. Second, each of the problem solving tasks explicitly included multiple cognitive operations from a set of four dimensions, that is, goal setting and progress monitoring, planning and self-organizing, acquiring and evaluating information, and making use of information, as defined



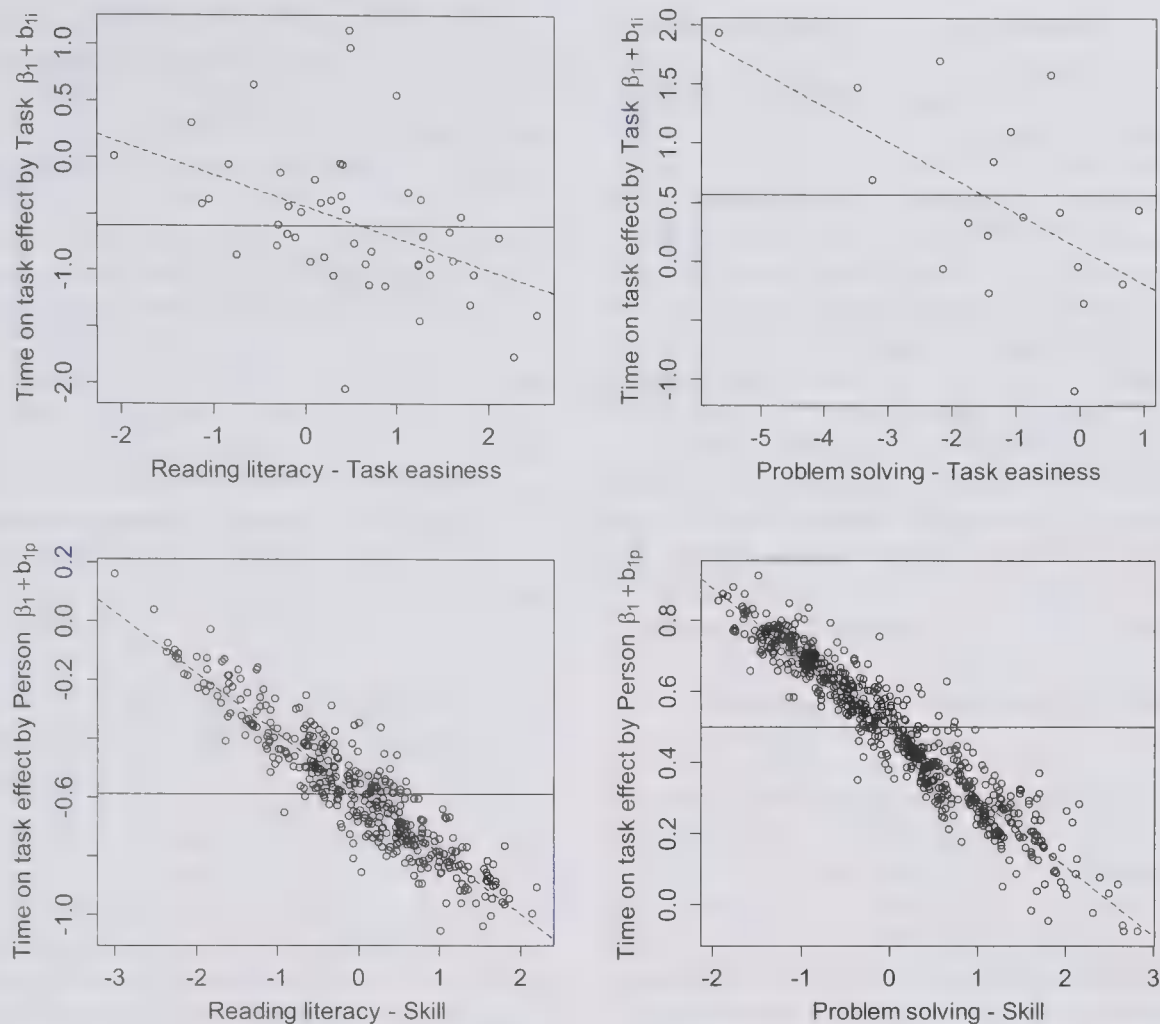


Figure 4. Upper row: Time on task effect by task for reading literacy (left panel) and problem solving in technology-rich environments (right panel). The solid line indicates the fixed time on task effect; the dots show how it is adjusted by task. For difficult tasks, the time on task effect gets more positive, whereas it gets more negative for easy tasks. Lower row: Time on task effect by person for reading literacy (left panel) and problem solving in technology-rich environments (right panel). The solid line indicates the fixed time on task effect; the dots show how it is adjusted by person. For less able individuals, the time on task effect gets more positive, whereas for able persons, it gets more negative.

by the PIAAC assessment framework (OECD, 2009b, p. 10). Given the constraints of a large-scale assessment, PIAAC only aimed at an overall indicator of problem solving. Our analyses would require a more fine-grained measure with a broad set of indicators for the various underlying cognitive operations. Although, for each task, one operation is assumed to be dominant, other operations might also be involved. For instance, the PIAAC framework maps the sample task “Job Search” to the cognitive operations of access and evaluating information as well as monitoring criteria for constraint satisfaction. There were only two more tasks that showed a comparable set of assumed cognitive operations, whereas in other tasks the requirement of accessing information was combined with a different additional demand, for example, communicating information. Thus, it was not possible to form subgroups with a sufficient number of tasks being homogeneous in the assumed composition of required cognitive operations.

### Time on Task Effect by Person (Hypothesis 3)

On the person level, we assumed that the effect of time on task varies across the individual skill level. To test Hypothesis 3, we

extended Model M0 to Model M3 by adding a random time on task effect,  $b_{1p}$ , representing the variation across individuals:  $\eta_{pi} = (\text{intercept } \beta_0) + (\text{individual skill } b_{0p}) + (\text{relative easiness } b_{0i}) + \beta_1 (\text{time on task } t_{pi}) + b_{1p} (\text{time on task } t_{pi})$ .

**Reading literacy.** For reading literacy, the variance of the by-person adjustment was  $\text{Var}(b_{1p}) = 0.14$ . Thus, for reading literacy, the time on task effect varied across persons. Most importantly, a correlation between the by-person time on task effect and by-person intercept of  $\text{Cor}(b_{0p}, b_{1p}) = -.65$  was estimated. That is, the overall negative time on task effect became stronger in able readers but was attenuated in poor readers. The bottom left panel in Figure 4 illustrates how the time on task effect adjusted by person linearly decreases in more able persons. To clarify whether the liberal Model M3 better fitted the data, we compared the nested Models M0 and M3. The model difference test revealed that Model M3 fitted the data significantly better than Model M0,  $\chi^2(2) = 15.09$ ,  $p < .01$ . To test whether the correlation parameter is required to improve model fit, that is, to test the significance of the correlation, Model M3 was compared with a restricted version (Model M3r) without the correlation between by-person time on task effect and intercept. The model difference test revealed that



Model M3 without restrictions was the better fitting model,  $\chi^2(1) = 12.85$ ,  $p < .01$ .

**Problem solving.** Similar results were obtained for problem solving. The variance of the by-person adjustment to the fixed effect of time on task was  $\text{Var}(b_{1p}) = 0.22$ . Thus, for problem solving in technology-rich environments, the time on task effect varied across persons. The correlation between the by-person adjustment of the time on task effect and the by-person intercept (individual skill) was again negative and substantial,  $\text{Cor}(b_{0p}, b_{1p}) = -.79$ . That is, the overall positive time on task effect became even stronger in poor problem solvers but was attenuated in able problem solvers (see the bottom right panel in Figure 4). The difference test comparing Model M3 including the random time on task effect with the baseline Model M0 was almost significant,  $\chi^2(2) = 5.98$ ,  $p = .05$ . Finally, comparing Model M3 with a restricted version (Model M3r) without a correlation between by-task time on task effect and intercept revealed that the correlation was significant,  $\chi^2(1) = 5.98$ ,  $p = .01$ .

### Integrated Model: Time on Task Effect by Task and Person

As assumed in Hypotheses 2 and 3, the previous results indicate that task difficulty and individual skill level have an influence on the strength and direction of the time on task effect. The final Model M4 integrates both the by-task and the by-person adjustments to the time on task effect. The results found for Models M1 and M3 were perfectly reproduced in the following Model M4:  $\eta_{pi} = (\text{intercept } \beta_0) + (\text{individual skill } b_{0p}) + (\text{relative easiness } b_{0i}) + \beta_1 (\text{time on task } t_{pi}) + b_{1i} (\text{time on task } t_{pi}) + b_{1p} (\text{time on task } t_{pi})$ .

**Reading literacy.** For reading literacy, the time on task effect was estimated to be  $\beta_1 = -0.69$  ( $z = -5.16$ ,  $p < .01$ ). The variance of the by-task adjustment to the time on task effect was  $\text{Var}(b_{1i}) = 0.64$ , and that of the by-person adjustment was  $\text{Var}(b_{1p}) = 0.23$ , that is, the time on task effect varied across both reading tasks and readers. Moreover, the time on task effect varied systematically in that the adjustments were linearly related to task easiness and individual skill level, respectively, as expected. The correlation between easiness of reading tasks and by-task adjustment was  $\text{Cor}(b_{0i}, b_{1i}) = -.52$ , and the correlation between individual skill and by-person adjustment was  $\text{Cor}(b_{0p}, b_{1p}) = -.78$ . The difference test showed that model M4 fit the data significantly better than model M0,  $\chi^2(4) = 106.14$ ,  $p < .001$ .

The curves in Figure 5 (upper panel) indicate how for a given reader and reading task the probability for a correct response depends on time on task. The range of the time on task axis represents the empirical range of time on task in the selected tasks. The slope of the curves resulted from adding up the time on task effect and the adjustments to the time on task effect by task and by person. When considering a proficient reader (skill level of  $b_{0p} = 1.61$ ) and an easy reading task (easiness of  $b_{0i} = 1.89$ ), that is, a reading situation of low demand, the unadjusted negative effect of  $-.69$  became much stronger, resulting in a negative time on task effect of  $-1.90$  (plus line). However, in a situation of high demand, where a difficult reading task (easiness of  $b_{0i} = -0.77$ ) was completed by a poor reader (skill level of  $b_{0p} = -1.79$ ), the curve's slope was no longer negative but even slightly positive, that is,  $0.55$  (triangle line). In situations of medium demand, that is, a poor reader completing an easy task or an able reader completing a difficult task, the curves' slopes are in-between.

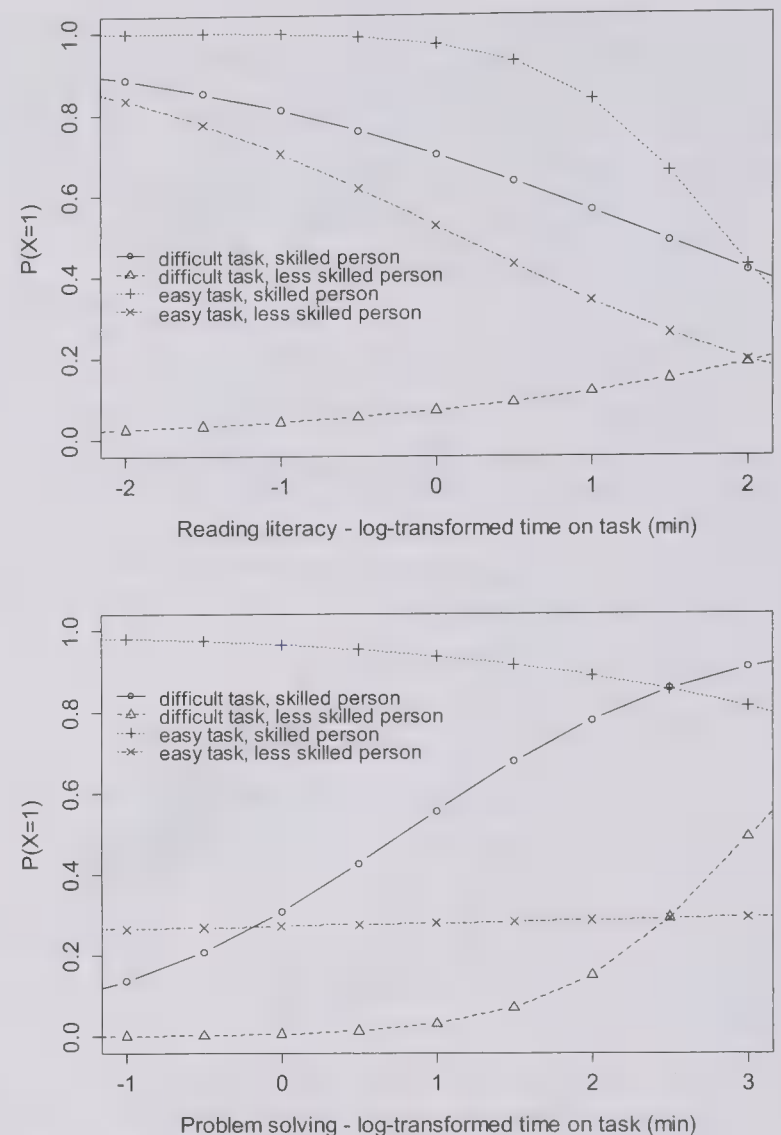


Figure 5. Time on task effect by task and skill level for reading literacy (upper panel) and problem solving in technology-rich environments (lower panel). For combinations of two tasks (easy vs. hard) with two persons (less able vs. able), the probability of obtaining a correct response is plotted as a function of time on task.

**Problem solving.** In the integrated model, a positive time on task effect of  $\beta_1 = 0.56$  ( $z = 2.26$ ,  $p = .02$ ) was obtained. The variance of the by-task adjustment to the time on task effect was  $\text{Var}(b_{1i}) = 0.89$ , and that of the by-person adjustment was  $\text{Var}(b_{1p}) = 0.11$ . The correlation between easiness of problem solving tasks and the by-task adjustment to the time on task effect was  $\text{Cor}(b_{0i}, b_{1i}) = -.63$ , and the correlation between individual skill level and the by-person adjustment to the time on task effect was  $\text{Cor}(b_{0p}, b_{1p}) = -.76$ . Again, the model comparison test indicated that model M4 fit the data significantly better than model M0,  $\chi^2(4) = 76.77$ ,  $p < .001$ .

The bottom panel in Figure 5 shows the probability of obtaining a correct response as a function of the time on task for two selected tasks completed by two selected persons. In a situation of low demand, that is, a proficient problem solver (skill level of  $b_{0p} = 2.63$ ) completing an easy task (easiness of  $b_{0i} = -0.67$ ), the time on task effect decreases dramatically and becomes even negative and was estimated as  $-0.62$  (+ line in Figure 5). However, in the situation of high demand where a difficult task (easiness of  $b_{0i} =$



−3.44) is completed by a poor problem solver (skill level of  $b_{op} = -1.66$ ), the positive time on task effect of .56 becomes much stronger and was estimated as 1.69 ( $\Delta$  line in Figure 5). If the demand is medium, that is, a less able person completes an easy task or an able person completes a difficult task, the curves' slopes are in-between.

Taken together, these results indicate that positive time on task effects are observed especially in highly demanding situations, where not-so-skilled readers or problem solvers are confronted with a difficult task. Presumably, they can partly compensate for task demands by allocating cognitive resources. If this interpretation holds true, differential time on task effects should be observable on a within-task level as well. Specifically, if it is the strategic allocation of processing time that drives a positive time on task effect in problem solving tasks and difficult reading tasks being encountered by poor readers, on a within-task level the positive time on task effect should be confined to the processing of task-relevant parts of the stimulus. We tested this hypothesis as a last step.

### Decomposing the Time on Task Effect at the Task Level (Hypothesis 4)

Using fine-grained time information extracted from log files, we decomposed the global time on task into several components that reflect particular steps of task solution. This was done at the task level for the problem solving task "Job Search," which required screening a search engine results page (see Figure 1, lower panel) and visiting multiple linked Web pages. Two of five Web pages in this task meet the criteria specified in the instruction and have to be bookmarked to obtain a correct response. In Hypothesis 4, spending more time on the two target pages was expected to indicate strategic behavior associated with a higher probability of successful task completion. In contrast, a negative effect was assumed for spending time on the search engine results page, which did not provide any hints about the target pages. For the time spent on nontarget pages, a negative effect was also expected.

First, logistic regression was used to predict the task success by time on task. The sample size for this analysis was 182. This analysis revealed a nonsignificant time on task effect of  $-0.29$  ( $z = -0.59, p = .55$ ). As a second step, task success was predicted by the time spent on the search engine results page, the time spent on the two relevant Web pages, and the time spent on the three irrelevant Web pages. The obtained effect for time spent on the relevant Web pages was positive and significant as expected,  $0.96$  ( $z = 2.53, p = .01$ ), that is, spending more time on the target pages for evaluating the accessed information and monitoring the multiple criteria for constraint satisfaction was associated with a higher probability of achieving a correct response. In contrast, for the time spent on the search engine results page, a significant negative effect of  $-1.78$  was revealed ( $z = -2.97, p < .01$ ). The time spent on irrelevant Web pages was not significantly related to task success (estimated effect of  $0.13, z = 0.23, p = .82$ ). As a measure of effect size, we computed Nagelkerke's  $R^2$ , which was .25, that is, about a quarter of the response variability could be explained by the component time predictors. This result pattern suggests that successful problem solvers quickly discarded the irrelevant search engine results page, whereas relevant pages meeting evaluation criteria were checked carefully. This pattern is fully

compatible with the view that positive time on task effects in difficult tasks are due to a strategic allocation of cognitive resources, as already suggested by the moderation of the time on task effect by domain, task difficulty, and skill level.

## Discussion

Computer-based assessment provides new possibilities to assess cognitive skills and underlying processes by measuring not only the outcome of a task but also behavioral process data that might be interpreted in terms of cognitive processes happening throughout task completion. This means that to some degree, data from computer-based assessments may be used to address research questions through means of process analysis that were previously confined to experimental research. This is of interest especially in combination with the rather large sample sizes obtained in educational assessments (compared to lab experiments). Thus, while there used to be a tradeoff—either go with small samples and deep process analysis or have large samples and test result data only—this tradeoff can be remedied to some degree by using process data from large-scale assessments.

The goal of this study was to investigate the effect of time on task on task success in reading literacy and problem solving in technology-rich environments and to test potential moderating variables. Our central hypothesis was that the relative degree of strategic versus routine cognitive processing as required by a task, as well as the test taker's acquired skill, determines the strength and direction of the time on task effect. Accordingly, our results revealed that the time on task effect was moderated by domain, task difficulty, and individual skill.

### Time on Task Effects in Reading Literacy

For reading literacy, overall, a negative time on task effect was found, that is, brief times on task were associated with correct responses, and taking more time apparently was not related to greater task success. Very slow respondents thus fail on the task. This observation especially concerns easy reading tasks as shown by the negative correlation between task easiness and the task-specific time on task effect, which means that for easy tasks the time on task effect was more negative than for difficult ones. To put it simply, in very easy tasks, the correct solution was either obtained quickly or never. In contrast, for difficult reading tasks, this association got weaker and in some instances was reversed. Taking individual differences in reading skill into account, these findings were consistently extended, that is, with increasing reading skill, the time on task effect got more negative, whereas it got weaker or even positive with decreasing reading skill. Thus, for poor readers completing hard reading tasks, time on task showed a positive effect, whereas for proficient readers working on easy tasks, a very strong negative effect was found. The latter result means that the few proficient readers who did not master the easy reading tasks took more time than the majority of proficient readers who were successful. In contrast, in a group of less skilled readers, this time difference between correct and incorrect answers in the same tasks was less pronounced, as shown by the weaker negative time on task effect.

The observed result pattern that incorrect responses are associated with longer times on task has consistently been found for



other untimed performance measures as well, for instance, general knowledge tasks (Ebel, 1953), matrices tasks (Hornke, 2000), figure series, number series, verbal analogy tasks (Beckmann, 2000), verbal memory tasks (Hornke, 2005), and discrimination tasks (for a review of reaction time research on this matter, see Luce, 1986). Hornke (2005) discussed how correct responses with short latencies are eye-catching. Incorrect responses in contrast may be preceded by an ongoing process of rumination and ultimately a switch to random guessing. This interpretation is consistent with our finding in that, especially for easy tasks, there is a strong negative time on task effect and also explains why, in easy reading tasks, generally skilled readers had a lower chance of getting the task correct when the response took longer. Similar effects were reported by Hornke (2000) and Beckmann (2000).

Across the cognitive operations required in reading tasks, there was no significant variation of the time on task effect. Thus, differences in the time on task effect across tasks cannot be ascribed to the presence or absence of specific cognitive operations as outlined in the PIAAC framework. In line with our findings on the dependency of the time on task effect on task difficulty, the clusters of access, integrate, and evaluate tasks are not very well distinguishable by their level of difficulty. Other task features than the cognitive operations are hence responsible for the variation of the time on task effect with task difficulty. If our cognitive interpretation of time on task effects holds, it might be worthwhile to look for task features that drive task difficulty and differential time on task effects. Identifying these features might further contribute to clarifying the PIAAC reading tasks' demands in cognitive terms and as such contribute to further advance the assessment framework. Therefore, as one future step, we intend to classify the PIAAC tasks, for instance, in terms of the transparency of the information, or the degree of complexity in making inferences (cf. OECD, 2009b). Task features such as these are not yet entirely covered by the aspects detailed by the PIAAC assessment framework.

### Time on Task Effects in Problem Solving

For problem solving, overall, a significant positive time on task effect was found: Long times on task were associated with correct answers and short times on task with wrong answers. Similar to reading, the time on task effect varied significantly across tasks. For easy tasks, it was weaker and around zero, whereas for difficult tasks, it became even more positive. This means that when dealing with challenging problems, spending more time was associated with higher probability of giving a correct response. Across individuals, poor problem solvers could benefit more from spending more time on a task than strong problem solvers. Although causal interpretations are not possible, this result suggests that poor problem solvers can compensate for their lack of general skill by putting in more effort when working on a particular task, especially when this task is hard to solve. Thus, the difference in time on task between correct and incorrect solutions was greater for weak problem solvers than for strong problem solvers, which is the reverse of the finding for reading.

The results on the time on task effect for reading literacy and problem solving show that the moderating role of task difficulty and person's skill are similar for both domains, even though the overall effect is very different. The time on task effect may become

similar between the two domains when considering the extreme cases in which a skilled person encounters an easy task or a less skilled person engages in a difficult task. In the first case, the resulting time on task effect is negative (even for problem solving), and in the second case, it is positive (even for reading literacy). Thus, across domains the strength and the direction of the time on task effects seem to be governed by skill and difficulty in the same way. Both high skill levels and easy tasks presumably are associated with a large proportion of cognitive component processes that are apt to automatization (in easy tasks) or in fact automatized (in skilled persons), bringing about a negative time on task effect. In contrast, low skill levels and difficult tasks presumably are associated with the need to engage in controlled and thus time-consuming cognitive activity to a large extent, bringing about a positive time on task effect.

Thus, on the one hand, problem solving and reading are conceived as involving different cognitive processes, and overall the relation of time on task to task success also clearly differs between the two domains. On the other hand, our results support the notion that combinations of tasks and persons form a continuum across the two domains ranging from automatic processing to controlled processing. Practicing a task may move a person-task combination to automatic processing. However, this is limited by the nature of the task. For instance, certain aspects of a problem solving task may become automated in skilled individuals, but not core aspects of problem solving, such as inducing rules or drawing conclusions.

Our interpretations of the time on task effect are further backed by the in-depth analysis of the time-taking behavior in the sample problem solving task "Job Search." This analysis was based on time data that was assumed to reflect different steps of task solution and presumably information processing. It revealed that only for time spent on steps that are necessarily needed to solve the tasks, that is, to visit and evaluate the target pages for multiple criteria, a positive time on task effect emerged, whereas for spending time on the noninformative search engine results page and the nontarget pages, negative or null effects were found. When spending time on the target pages, the problem solver is assumed to deal with the part of the problem space that enables one to move step by step to the knowledge state that includes the solution (Simon & Newell, 1971) or to integrate relevant information, rather than identifying various other aspects of the problem (Wirth & Leutner, 2008). Thus, this finding supported our hypothesis that the positive time on task effect in problem solving tasks reflects the need for and the benefit from devoting time to strategic and controlled cognitive processing. This interpretation suggests that task success could depend on the time spent on relevant pages (however, time on task as well as task success might also be driven by a common cause such as motivation). The negative effect of time spent on the search engine results page may indicate the strategy to select Web pages based on the limited information provided there. Although this approach could in principal be useful to filter search results, in the given task the results page did not indicate whether search criteria would be met or not. Thus, lingering on a page that could not contribute to solving the task was in fact detrimental to succeeding.

### Time on Task and a Dual Processing Framework

We derived our hypotheses on differential time on task effects both between and within domains by means of applying a dual



processing framework (cf. Fitts & Posner, 1967; Schneider & Chein, 2003; Schneider & Shiffrin, 1977) to reading and problem solving tasks used in the PIAAC study. The hypotheses thus derived were confirmed; hence, our results are consistent with the notion that positive time on task effects reflect the strategic allocation of cognitive resources, whereas negative time on task effects reflect the degree of automatization. Although the findings are entirely consistent with the predictions derived from such a framework and further backed by analyses on a within-task level, this interpretation has to remain somewhat speculative for the time being. The information that can be gained from large-scale computer-based assessments (although providing much more information than traditional paper-and-pencil based assessments) is still limited. Usually, the information stored in log files is ambiguous as to its interpretation in cognitive terms. In this article, we have assumed that taking more time on more difficult tasks indicates engaged cognitive processing. Other interpretations of the pattern of results are yet conceivable. For instance, it might be the case that time on task effects also reflect differences in motivation, that is, test takers not only take *more time* to think about a task but also think *harder*—resulting in a confounding between depth of processing and time taken. Related to that, Guthrie et al. (2004) considered time on task as an indicator of engagement, which means to read a text attentively, concentrating on the meaning, and with sustained cognitive effort (see also Kupiainen et al., 2014). Issues such as these can only be resolved by combining the analysis of large-scale process data with research tools allowing for an even more fine-grained analysis of cognitive processes, such as eye movements or think-aloud techniques (see Rouet & Passerault, 1999). As a consequence, we aim at corroborating our results through experimental studies that combine actual large-scale testing materials and still more fine-grained assessments of cognitive processes in the future.

## Limitations

In the present study, test takers were free to adapt their speed–accuracy compromise both within and between tasks, which has consequences on the interpretation of the obtained results. As the speed level of test takers was not controlled, the obtained variation in the association between time on task and task success across tasks may be due to different task difficulties as claimed in Hypothesis 2 or due to within-person differences in the selected speed level across tasks. However, the latter explanation does not seem plausible as there is empirical evidence for the assumption of stationarity of speed when completing power tests (cf., e.g., Goldhammer & Klein Entink, 2011; Klein Entink et al., 2009). Stationarity of speed is also implied by the fixed level of accuracy which is a standard assumption in item response models (cf. van der Linden, 2007).

As we did not manipulate the speed level of test takers experimentally, we cannot conclude that the predictor time on task has any causal effect on task success, which, however, is suggested by the positive time on task effect in those tasks requiring a higher level of controlled processing. In contrast, in tasks that can be completed more automatically and for which a negative effect was revealed, time on task should rather be conceived of as an indicator of competence in addition to the task result.

As another limitation, the sample size of the present study and the number of responses per task, respectively, were quite limited for testing measurement models. Therefore, future research should aim at replicating the findings based on greater samples, for instance, from the PIAAC main study. Another important replication goal would be to investigate whether results on the time on task effect are comparable across countries.

In PIAAC the construct of problem solving in technology-rich environments was newly developed as was the measurement procedure. Thus, future research will have to provide more information about this assessment's validity and its predictive power. Moreover, the relation of problem solving in PIAAC to other problem solving measures and their theoretical underpinnings requires further clarification. There are several conceptual commonalities, for example, representing the difference between a current state and a goal state, defining a series of subgoals, and applying related nonroutine cognitive and behavioral operations to transform the given state into the targeted state, including progress monitoring. However, there are also remarkable differences. For instance, the construct of complex problem solving (cf. Funke & Frensch, 2007) assumes systems where complexity is defined by the number of elements and the relations among them. The problem solving process is comprised of the acquisition of knowledge by means of exploration and the application of the obtained knowledge. Although acquiring knowledge or information is also a key aspect of problem solving as defined in PIAAC, acquired knowledge in a complex problem solving task represents the explored system of elements and relations itself. In contrast, in PIAAC problem solving, the explored system is just the medium carrying the information that is required to solve the task. However, an unfamiliar computer environment and unknown functionality would turn the problem solving in technology-rich environments task into a complex problem solving task (for technical problem solving, see, e.g., Baumert, Evans, & Geiser, 1998). Regarding our findings on problem solving as proposed by PIAAC, future research needs to show whether the pattern of results holds true also for other conceptions of problem solving that, for instance, are anchored in cognitive theory (see, e.g., Fischer, Greiff, & Funke, 2012) or used in other large-scale assessments such as the Programme for International Student Assessment (PISA; cf. Greiff, Holt, & Funke, 2013).

## Educational Implications

The present study frames the meaning of time in information-processing tasks by referring to models of skill acquisition and related individual differences. Therefore, although the analyses are based on assessment tasks, our results allow for some tentative conclusions on educational procedures in reading and problem solving instruction. Our results indicate that for learning and applying higher level cognitive skills, required component skills should be well routinized. If there is no established routine processing, for instance, when a poor reader encounters difficult reading tasks, information processing needs to rely on strategic processing as indicated by the reversed positive effect of time on task on task success. This means that for poor readers to be successful, they need to switch to compensatory behaviors, that is, reducing reading rate, looking back in text, reading aloud, and pausing, and/or compensatory strategies, that is, shifting attention



to lower level requirements and rereading text, to cope with their deficits. Following Walczyk's (2000) compensatory-encoding model, "With enough time, any text can be vanquished!" (p. 565).

From an instructional point of view this means that becoming a good reader or problem solver requires the development of self-regulatory and metacognitive skills necessary to know when an effortful, controlled processing mode is to be employed. In the controlled processing mode, appropriate compensatory mechanisms can be initiated that have been learned and incorporated before. This might, for example, mean that in the face of reading comprehension difficulties, a part of a text is reread or that in problem solving, time is taken to focus attention on relevant subgoals.

As the individual time on task effect is assumed to reflect the way of processing information, it may help to further describe the individual performance level and to identify instructional needs. As suggested in Figure 4 (bottom left panel), average readers show a great variation in the time on task effect, suggesting various levels of automaticity of component skills. Moreover, the in-depth investigation of temporal patterns in highly interactive tasks such as problem solving tasks can point to deficits in the information-processing strategy (cf. Zoanetti, 2010). For instance, if, in the "Job Search" task, log file data would reveal that a problem solver spends much time both on nontarget pages and target pages, this pattern would suggest that the problem solver cannot process disconfirming information efficiently to quickly discard a nontarget page.

From an educational measurement perspective, the present study suggests that the meaning of time on task is not uniform. Thus, when collecting time information across tasks and individuals that are heterogeneous in difficulty and skill level, respectively, the role of time and its interpretation may differ. Regarding item response models including time as a regressor, van der Linden (2007) argued that time can only be interpreted uniformly as an indicator of speed if the tasks do not differ substantially in the amount of labor. In the present study where tasks differ considerably in the amount of information processing and problem solving, we take the different interpretations of time on task into account by letting its effect vary across tasks (random effect).

All in all, the analyses and results reported here illustrate the potentials that lie in exploiting time on task, or fractions of it, that become available through computer-based assessments. They do however also clarify that any process measure must be cautiously interpreted, at least by taking a closer look at the particular tasks and their demands. Regarding the two constructs studied here, reading literacy and problem solving in technology-rich environments, our study proves them to be quite different in terms of cognitive processing. Skill, task difficulty, and time on task do interact in different ways. As Wirth and Klieme (2003) have shown based on student assessment in a German national extension to PISA, problem solving tests, especially computer-based problem solving tests, add to the traditional set of literacy dimensions. In structural models, problem solving skills can be clearly distinguished from traditional abilities such as reasoning (cf. Wüstenberg, Greiff, & Funke, 2012). These structural analyses and our in-depth analyses of processing time provide evidence that problem solving skills have to be separated from traditional educational outcomes such as reading literacy. Problem solving is one of the most prominent examples of cross-curricular, nonroutine,

dynamic 21st century skills that are currently aimed at as educational goals and covered in large-scale surveys. Claims that these new skills are different from traditional outcomes have mainly been supported by pragmatic or philosophical arguments. Now, we see that even in terms of cognitive processing and time allocation, there is a difference between reading literacy and problem solving skills.

## References

- Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin*, 102, 3–27. doi:10.1037/0033-2909.102.1.3
- Ackerman, P. L., & Cianciolo, A. T. (2000). Cognitive, perceptual speed, and psychomotor determinants of individual differences during skill acquisition. *Journal of Experimental Psychology: Applied*, 6, 259–290. doi:10.1037/1076-898X.6.4.259
- Anderson, J. R. (1992). Automaticity and the ACT\* theory. *American Journal of Psychology*, 105, 165–180. doi:10.2307/1423026
- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390–412. doi:10.1016/j.jml.2007.12.005
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133, 283–316. doi:10.1037/0096-3445.133.2.283
- Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using Eigen and S4 (R Package Version 0.999999–0) [Computer software]. Retrieved from <http://CRAN.R-project.org/package=lme4>
- Baumert, J., Evans, R. H., & Geiser, H. (1998). Technical problem solving among 10-year-old students as related to science achievement, out-of-school experience, domain-specific control beliefs, and attribution patterns. *Journal of Research in Science Teaching*, 35, 987–1013. doi:10.1002/(SICI)1098-2736(199811)35:9<987::AID-TEA3>3.0.CO;2-P
- Beckmann, J. F. (2000). Differentielle Latenzzeiteffekte [Differential effects of latencies]. *Diagnostica*, 46, 124–129. doi:10.1026/0012-1924.46.3.124
- Best, R., Rowe, M., Ozuru, Y., & McNamara, D. (2005). Deep-level comprehension of science texts: The role of the reader and the text. *Topics in Language Disorders*, 25, 65–83. doi:10.1097/00011363-200501000-00007
- Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology and Evolution*, 24, 127–135. doi:10.1016/j.tree.2008.10.008
- Carlson, R. A., Khoo, B. H., Yaure, R. G., & Schneider, W. (1990). Acquisition of a problem-solving skill: Levels of organization and use of working memory. *Journal of Experimental Psychology: General*, 119, 193–214.
- Carlson, R. A., Sullivan, M. A., & Schneider, W. (1989). Component fluency in a problem-solving context. *Human Factors*, 31, 489–502. doi:10.1177/001872088903100501
- Carver, R. P. (1992). Reading rate: Theory, research and practical implications. *Journal of Reading*, 36, 84–95.
- De Boeck, P. (2008). Random item IRT models. *Psychometrika*, 73, 533–559. doi:10.1007/s11336-008-9092-x
- De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, 39, 1–28.



- De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer. doi:10.1007/978-1-4757-3990-9
- Dodonova, Y. A., & Dodonov, Y. S. (2013). Faster on easy items, more accurate on difficult ones: Cognitive ability and performance on a task of varying difficulty. *Intelligence*, 41, 1–10. doi:10.1016/j.intell.2012.10.003
- Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software*, 20, 1–18.
- Ebel, R. (1953). The use of item response time measurements in the construction of educational achievement tests. *Educational and Psychological Measurement*, 13, 391–401. doi:10.1177/001316445301300303
- Fischer, A., Greiff, S., & Funke, J. (2012). The process of solving complex problems. *Journal of Problem Solving*, 4, 19–41. doi:10.7771/1932-6246.1118
- Fitts, P. M., & Posner, M. I. (1967). *Learning and skilled performance in human performance*. Belmont, CA: Brooks/Cole.
- Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking & Reasoning*, 7, 69–89. doi:10.1080/13546780042000046
- Funke, J., & Frensch, P. A. (2007). Complex problem solving: The European perspective—10 years after. In D. H. Jonassen (Ed.), *Learning to solve complex scientific problems* (pp. 25–47). New York, NY: Erlbaum.
- Gelman, A. (2005). Analysis of variance? Why it is more important than ever. *Annals of Statistics*, 33, 1–53. doi:10.1214/009053604000001048
- Goldhammer, F., & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence*, 39, 108–119. doi:10.1016/j.intell.2011.02.001
- Goldhammer, F., Naumann, J., & Keßel, Y. (2013). Assessing individual differences in basic computer skills: Psychometric characteristics of an interactive performance measure. *European Journal of Psychological Assessment*. Advance online publication. doi:10.1027/1015-5759/a000153
- Gräsel, C., Fischer, F., & Mandl, H. (2000). The use of additional information in problem-oriented learning environments. *Learning Environments Research*, 3, 287–305. doi:10.1023/A:1011421732004
- Greiff, S., Holt, D. V., & Funke, J. (2013). Perspectives on problem solving in educational assessment: Analytical, interactive, and collaborative problem solving. *Journal of Problem Solving*, 5, 71–91. doi:10.7771/1932-6246.1153
- Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational settings—Something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology*, 105, 364–379. doi:10.1037/a0031856
- Guthrie, J. T., Wigfield, A., Barbosa, P., Perencevich, K. C., Taboada, A., Davis, M. H., . . . Tonks, S. (2004). Increasing reading comprehension and engagement through Concept-Oriented Reading Instruction. *Journal of Educational Psychology*, 96, 403–423. doi:10.1037/0022-0663.96.3.403
- Hornke, L. F. (2000). Item response times in computerized adaptive testing. *Psicológica*, 21, 175–189.
- Hornke, L. F. (2005). Response time in computer-aided testing: A “verbal memory” test for routes and maps. *Psychology Science*, 47, 280–293.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, England: Cambridge University Press.
- Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, 74, 21–48. doi:10.1007/s11336-008-9075-y
- Klieme, E. (2004). Assessment of cross-curricular problem-solving competencies. In J. H. Moskowitz & M. Stephens (Eds.), *Comparing learning outcomes: International assessments and education policy* (pp. 81–107). London, England: Routledge.
- Krajewski, K., & Schneider, W. (2009). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: Findings from a four-year longitudinal study. *Learning and Instruction*, 19, 513–526. doi:10.1016/j.learninstruc.2008.10.002
- Kupiainen, S., Vainikainen, M.-P., Marjanen, J., & Hautamäki, J. (2014). The role of time on task in computer-based assessment of cross-curricular skills. *Journal of Educational Psychology*, 106, 627–638. doi:10.1037/a0035507
- Landerl, K., & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology*, 100, 150–161. doi:10.1037/0022-0663.100.1.150
- Luce, R. D. (1986). *Response times: Their roles in inferring elementary mental organization*. Oxford, England: Oxford University Press.
- Mayer, R. E. (1992). *Thinking, problem solving, cognition* (2nd ed.). New York, NY: Freeman.
- Mayer, R. E. (1994). Problem solving, teaching and testing for. In T. Husén & T. N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed., Vol. 8, pp. 4728–4731). Oxford, England: Pergamon Press.
- McKeown, M. G., Beck, I. L., & Blake, R. G. K. (2009). Rethinking reading comprehension instruction: A comparison of instruction for strategies and content approaches. *Reading Research Quarterly*, 44, 218–253. doi:10.1598/RRQ.44.3.1
- Morrell, C. H., Pearson, J. D., & Brant, L. J. (1997). Linear transformations of linear mixed-effects models. *American Statistician*, 51, 338–343. doi:10.1080/00031305.1997.10474409
- Naumann, J., Richter, T., Christmann, U., & Groeben, N. (2008). Working memory capacity and reading skill moderate the effectiveness of strategy training in learning from hypertext. *Learning and Individual Differences*, 18, 197–213. doi:10.1016/j.lindif.2007.08.007
- Naumann, J., Richter, T., Flender, J., Christmann, U., & Groeben, N. (2007). Signaling in expository hypertexts compensates for deficits in reading skill. *Journal of Educational Psychology*, 99, 791–807. doi:10.1037/0022-0663.99.4.791
- Neubauer, A. C. (1990). Speed of information processing in the Hick paradigm and response latencies in a psychometric intelligence test. *Personality and Individual Differences*, 11, 147–152. doi:10.1016/0191-8869(90)90007-E
- Nicolson, R. I., & Fawcett, A. J. (1994). Reaction times and dyslexia. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 47(A), 29–48. doi:10.1080/14640749408401142
- Organisation for Economic Co-Operation and Development. (2009a). *PIAAC literacy: A conceptual framework* (OECD Education Working Paper No. 34). Paris, France: Author.
- Organisation for Economic Co-Operation and Development. (2009b). *PIAAC problem solving in technology-rich environments: A conceptual framework* (OECD Education Working Paper No. 36). Paris, France: Author.
- Organisation for Economic Co-Operation and Development. (2011). *PISA 2009 results: Vol. VI. Digital technologies and performance*. Paris, France: Author.
- Organisation for Economic Co-Operation and Development. (2013). *OECD skills outlook 2013: First results from the Survey of Adult Skills*. doi:10.1787/9789264204256-en
- Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 11, 357–383. doi:10.1080/10888430701530730



- Puntambekar, S., & Stylianou, A. (2005). Designing navigation support in hypertext systems based on navigation patterns. *Instructional Science*, 33, 451–481. doi:10.1007/s11251-005-1276-5
- R Core Team. (2012). R: A language and environment for statistical computing [Computer software]. Retrieved from <http://www.R-project.org/>
- Richter, T., Isberner, M.-B., Naumann, J., & Kutzner, Y. (2012). Prozessbezogene Diagnostik von Lesefähigkeiten bei Grundschulkindern [Process-based measurement of reading skills in primary school children]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology*, 26, 313–331. doi:10.1024/1010-0652/a000079
- Richter, T., Isberner, M.-B., Naumann, J., & Neeb, Y. (2013). Lexical quality and reading comprehension in primary school children. *Scientific Studies of Reading*. Advance online publication. doi:10.1080/10888438.2013.764879
- Richter, T., Naumann, J., Brunner, M., & Christmann, U. (2005). Strategische Verarbeitung beim Lernen mit Text und Hypertext [Strategic processing in learning from text and hypertext]. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology*, 19, 5–22. doi:10.1024/1010-0652.19.12.5
- Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). New York, NY: Springer. doi:10.1007/978-1-4757-2691-6\_11
- Rouet, J.-F., & Passerault, J.-M. (1999). Analyzing learner-hypermedia interaction: An overview of on-line methods. *Instructional Science*, 27, 201–219. doi:10.1007/BF00897319
- Scheuermann, F., & Björnsson, J. (2009). *The transition to computer-based assessment*. Luxembourg: Office for Official Publications of the European Communities.
- Schleicher, A. (2008). PIAAC: A new strategy for assessing adult competencies. *International Review of Education*, 54, 627–650. doi:10.1007/s11159-008-9105-0
- Schneider, W., & Chein, J. M. (2003). Controlled & automatic processing: Behavior, theory, and biological mechanisms. *Cognitive Science*, 27, 525–559. doi:10.1016/S0364-0213(03)00011-9
- Schneider, W., & Fisk, A. D. (1983). Attentional theory and mechanisms for skilled performance. In R. A. Magill (Ed.), *Memory and control of action* (pp. 119–143). New York, NY: North-Holland. doi:10.1016/S0166-4115(08)61989-5
- Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, 84, 1–66. doi:10.1037/0033-295X.84.1.1
- Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, 84, 127–190. doi:10.1037/0033-295X.84.2.127
- Simon, H. A., & Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American Psychologist*, 26, 145–159. doi:10.1037/h0030806
- Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. S. Haladyna (Eds.), *Handbook of test development* (pp. 329–347). Mahwah, NJ: Erlbaum.
- Sullivan, S., Gnesdilow, D., & Puntambekar, S. (2011). Navigation behaviors and strategies used by middle school students to learn from a science hypertext. *Journal of Educational Multimedia and Hypermedia*, 20, 387–423.
- van den Broek, P. W., & Espin, C. A. (2012). Connecting cognitive theory and assessment: Measuring individual differences in reading comprehension. *School Psychology Review*, 41, 315–325.
- van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer.
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72, 287–308. doi:10.1007/s11336-006-1478-z
- van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, 46, 247–272. doi:10.1111/j.1745-3984.2009.00080.x
- van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for speededness in computerized adaptive testing. *Applied Psychological Measurement*, 23, 195–210. doi:10.1177/01466219922031329
- Walczyk, J. (2000). The interplay between automatic and control processes in reading. *Reading Research Quarterly*, 35, 554–566. doi:10.1598/RRQ.35.4.7
- Wickelgren, W. A. (1977). Speed–accuracy tradeoff and information processing dynamics. *Acta Psychologica*, 41, 67–85. doi:10.1016/0001-6918(77)90012-9
- Wirth, J., & Klieme, E. (2003). Computer-based assessment of problem solving competence. *Assessment in Education: Principles, Policy & Practice*, 10, 329–345. doi:10.1080/0969594032000148172
- Wirth, J., & Leutner, D. (2008). Self-regulated learning as a competence: Implications of theoretical models for assessment methods. *Zeitschrift für Psychologie/Journal of Psychology*, 216, 102–110. doi:10.1027/0044-3409.216.2.102
- Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving: More than reasoning? *Intelligence*, 40, 1–14. doi:10.1016/j.intell.2011.11.003
- Zoanetti, N. (2010). Interactive computer based assessment tasks: How problem-solving process data can inform instruction. *Australasian Journal of Educational Technology*, 26, 585–606.

Received March 1, 2013

Revision received July 1, 2013

Accepted July 5, 2013 ■



# The Role of Time on Task in Computer-Based Low-Stakes Assessment of Cross-Curricular Skills

Sirkku Kupiainen, Mari-Pauliina Vainikainen, Jukka Marjanen, and Jarkko Hautamäki  
University of Helsinki

The role of time on task (TOT) for students' attainment in a low-stakes assessment of cross-curricular skills was examined using the log data collected in the computer-based assessment (CBA). Two structural equation models were compared: Model 1, in which students' test scores were explained by grade point average (GPA) together with mastery and detrimental motivational attitudes, and Model 2, in which TOT was added to the model to mediate the effects of GPA and the 2 motivational constructs. Fitting the models to nationally representative data of 4,249 Finnish 9th graders ( $M_{\text{age}} = 15.92$  years) confirmed the hypothesis that investment of time plays a key role in explaining test scores in low-stakes assessment even when prior ability (GPA) is taken into account. It was also confirmed that the effects of the detrimental attitudes on students' attainment were mediated by TOT. The study makes an important contribution to research regarding the role of motivational attitudes and TOT in low-stakes assessment, which is vital for the use of the assessment results in national and international benchmarking. It is concluded that log data provide a functional way to investigate time investment in CBA as an indicator for students' effort, yielding relevant implications for educational psychologists.

**Keywords:** time on task, low-stakes assessment, log data, motivational attitudes, cross-curricular skills

Ever since the 1990s, there has been a growing interest in the wider cognitive and affective goals of education. These cross-curricular skills are believed to indicate readiness for new learning and for successful adaptation to the rapidly changing demands of the future, and they collectively represent one of the reasons why large-scale low-stakes assessments are now at the forefront of the national and international education scene. The most prominent example of such a low-stakes assessment that regularly administers tests of cross-curricular skills is the Organisation for Economic Co-operation and Development's Programme for International Student Assessment. Also, many curricular assessments, such as the International Association for the Evaluation of Educational Achievement's Progress in International Reading Literacy Study and Trends in International Mathematics and Science Study, are presented to students as low-stakes assessments. The high policy visibility of these studies has led to the results being used for benchmarking at the international and the country levels (e.g., America Achieves, 2013; Grek, 2009). Yet, the low stakes have been shown to lead to reduced validity and reliability when students are not putting their best effort into the assessment because of the lack of personal consequences. They have also led to underestimated norms in later high-stakes tests on the basis of the results (Barry, Horst, Finney, Brown, & Kopp, 2010; Wise, 2006).

The problem of low stakes has been countered by adding to the assessments self-report questions regarding the effort students have invested in the assessment, but the reliability of self-reported effort has been shown to be unreliable because of untruthful responding and not accounting for change in effort along an entire test session (cf. Wise & Kong, 2005). Another approach, available in computer-based assessment (CBA), has been to use log data to look at the time students invest in assessment tasks (e.g., Schnipke & Scrams, 1997; Wise & Kong, 2005), and advances made in CBA during the past decades look promising in this respect (Greiff et al., 2013; Wang, Jiao, Young, Brooks, & Olson, 2008). A key research question in CBA has centered on response time (RT) as an indicator of student effort, which is a crucial prerequisite for the reliability and validity of the assessment results (Lee & Chen, 2011).

When interpreted as a time-related indicator for effort, RT can be seen to relate to Carroll's (1963) notion of time on task (TOT) in learning. In his model, Carroll considered learning to be determined by the ratio of the time needed for learning and the time spent on learning (see also Bloom, 1980; Karweit, 1982). Ever since, TOT has featured regularly in meta-analyses of factors pertaining to learning and to school achievement (Hattie, 2005; Scheerens & Bosker, 1997). Yet, unlike the literature on RT, Carroll did not refer directly to motivational factors in his model but used the term *engagement* as a conative term, referring to the act of being engaged.

Drawing on the two research traditions of TOT and RT and using the log data afforded by CBA, in this article, we investigate the role time plays in students' attainment in a cross-curricular learning-to-learn (LTL) assessment. In the Finnish framework, LTL is conceptualized as the interplay of an individual's cognitive competence and motivational and affective characteristics, aroused in a learning situation. Accordingly, LTL is assessed with a test

---

This article was published Online First February 17, 2014.

Sirkku Kupiainen, Mari-Pauliina Vainikainen, Jukka Marjanen, and Jarkko Hautamäki, Centre for Educational Assessment, University of Helsinki, Helsinki, Finland.

Correspondence concerning this article should be addressed to Mari-Pauliina Vainikainen, P.O. Box 9, 00014 University of Helsinki, Finland. E-mail: mari-pauliina.vainikainen@helsinki.fi



comprising cognitive tasks and a questionnaire with multiple attitudinal scales (cf. Hautamäki et al., 2002, 2006). The present article focuses on the role of TOT in the combination of students' motivational attitudes as disclosed in the questionnaire and their attainment in the LTL test, taking into account prior ability. In this study, the term *prior ability* is used to refer to students' cognitive competence as captured in school achievement (see Atkinson & Geiser, 2009; Cattell, 1987, pp. 139–145), indicated by their grade point average (GPA), whereas the term *test attainment* refers to performance in the assessment (i.e., the test score). In this, the study makes a fresh contribution to current literature regarding the reliability of low-stakes cross-curricular assessments for making psychological grounded recommendations for changing educational systems (Olson, 2003).

Carroll's concept of TOT has been chosen to reflect the cross-curricular character of the assessment tasks. Unlike curricular assessment, the tasks require not only the application of the recently learned knowledge and procedures but the assimilation of the novel rules of the tasks and their application in the ensuing items. Accordingly, time has been operationalized as the time spent on a whole task instead of using item-specific RTs, bringing the study closer to Carroll's TOT. Yet, by linking time use to motivational attitudes, the study relates to recent research on RT as an indicator for effort and, hence, also makes a contribution to literature in this field.

## TOT

In his model, Carroll considers students' aptitude for the task, their ability to understand instruction, and the quality of instruction as the main constituents of the time needed for learning (Carroll, 1963, 1989, p. 26), whereas the time allocated (outer constraint) and the time a student is ready to spend (inner constraint) are the two constituents of the time spent on learning.<sup>1</sup> The time needed and the time spent, together with the quality of instruction, determine learning as the outcome. In her review, however, Karweit (1982) pointed out that much of the research on TOT concentrated on just the time spent on learning, ignoring the role of students' ability to apply instruction as the component defining the time needed for learning. Hence, when reanalyzing the reviewed studies while taking into account students' prior ability as a measure of the time needed, the role of TOT for learning diminished to a third of the effect reported in the studies not making this distinction (Karweit, 1982; see also Gettinger, 1985; Karweit & Slavin, 1981).

Compared with earlier observation-based studies, the log data of CBA provide a relatively accurate measure of the time students spend engaged in learning or on assessment tasks. This allows relating TOT more rigorously with not only the results of the assessment but also with other data collected concurrently. Accordingly, whereas earlier research on TOT mainly concentrated on just the interrelations between prior ability, TOT, and learning, in this article, we extend the focus to the role of students' motivational attitudes in their readiness to exert themselves in the tasks. We bring a new point of view to TOT research, still considering the time needed and the time spent on learning but moving it closer to the interpretation of RT as an indicator for effort. In the present study, the object of prediction is students' attainment in the low-stakes LTL reasoning tasks, and, on the basis of Carroll (1963, 1989), it is hypothesized that TOT acts as a mediating factor for

the effect of both prior ability (in this case, earlier school achievement as indicated by GPA) and students' motivational attitudes as determinants of the time students are ready to spend on the assessment tasks (inner constraint of time).

## RT

Whereas TOT is generally used to refer to the time spent on learning in class or across school days, RT is used to indicate the time it takes a student to answer one specific item in an assessment (e.g., Lee & Chen, 2011; Schnipke & Scrams, 2002; Wise & Kong, 2005). Much of the research on RT focuses on the comparison between high-stakes testing and low-stakes assessment, using RT as an indicator for effort. The expectation is that when test results have important personal consequences for the students (high stakes), they will put more effort into the test. When stakes are low at the personal level (low stakes), students are expected to balance test taking with other interests (e.g., trying to avoid mental exertion), leading to reduced effort. This is seen to affect the validity and the reliability of the results (Kong, Wise, & Bhola, 2007; Wise & DeMars, 2005; Wise & Kong, 2005). Accordingly, RT research provides empirical evidence on the relation of some motivational attitudes to students' use of time in an assessment situation, providing the link between RT and TOT elaborated on in the present study.

Building on Schnipke and Scrams's (1997) notion of solution versus rapid guessing behavior in speeded tests, Wise and Kong (2005) introduced the concept of *response time effort* (RTE) to describe the proportion of test items for which the examinees exhibit one or the other of the two behaviors. Of interest for the present study is that they found RTE was not related to academic achievement as indicated by Scholastic Aptitude Test scores. Additionally, Wise and DeMars found no relation—and no difference—between students' prior ability and their self-reported motivation in high-stakes and low-stakes tests. Instead, they found a significant difference in test attainment between motivated and unmotivated examinees, leading them to suggest the use of motivation filtering when striving for reliable proficiency estimates (Sundre & Wise, 2003; Wise & DeMars, 2008).

Alternatively, Goldhammer et al. (2014) showed in their study that the time students spend on successfully completing assessment tasks differs according to item difficulty (positive correlation), student ability (negative correlation), and domain. Unlike in reading literacy, in problem solving, the relation between TOT and task success was positive even for the easier tasks, indicating an additional difficulty inherent to tasks for which students do not have a preformed formula to use and for which effortful processing and, thus, a longer RT is needed.

Research regarding RT supports the understanding that TOT acts as a mediating factor between students' attainment in an assessment, their prior ability, and their motivational attitudes, even if neither TOT nor RT corresponds in their classical sense

<sup>1</sup> Depending on what is considered to be a task, there may be alternative definitions of TOT. For instance, in this special section, Goldhammer et al. (2014) used the term *time on task* to refer to the time needed to respond to a single question or problem, which is comparable to our notion of RT. We use the term *time on task* to refer to the time needed to complete a whole LTL reasoning task with instructions, examples, and multiple items to be solved.



fully to the way TOT has been operationalized in the present study. Hence, it reflects the understanding of RT as (partially) an indicator for effort and the results regarding the relative role of cognitive competence and various motivational factors in explaining achievement (e.g., Spinath, Spinath, Harlaar, & Plomin, 2006).

### Present Study

The study springs from a longstanding research project on LTL as one of the cross-curricular competencies the school is expected to foster in students (Csapó, 2007; Hautamäki et al., 2006; Hoskins & Fredriksson, 2008). The developed instrument focuses on the cognitive and affective factors salient for new learning and accessible in school-based assessment (Hautamäki et al., 2002, p. 5). The cognitive tasks are related to curricular content but not directly repeating it to require higher level application of general cognitive ability or reasoning in addition to the curricular knowledge and skills learned at school (Hautamäki et al., 2002; see also Adey, Csapó, Demetriou, Hautamäki, & Schayer, 2007, for the influence of education on general cognitive ability). For the self-report questionnaire, which comprises scales for diverse motivational and affective constructs shown to bear on present and future learning, three scales measuring attitudes positively related to learning and three scales measuring attitudes detrimental for learning were chosen.

To address students' investment of time in the assessment, we chose to use Carroll's (1963) learning model to emphasize that in cross-curricular assessment students face the same constraints of time, motivation, and prior abilities as in all learning. However, the focus of the study is on students' engagement and attainment in the assessment tasks instead of on their school achievement as in much of TOT research (Karweit, 1982). Therefore, we use students' prior school achievement (GPA) as an indicator of ability for the next stage of learning (Atkinson & Geiser, 2009; Gustafsson & Carlstedt, 2006; Thorsen & Cliffordson, 2012) instead of seeing school achievement as just the end state (grades given in an end-of-year report). This reflects Snow's (1990) model for educational assessment, in which he posits the relations between reasoning skills and school achievement as a cyclical transition from goal-relevant initial states to desired end states, which will, in turn, be the initial states for a next cycle of learning.

To reflect the double focus of the study combining Carroll's (1963, 1989) concept of TOT in terms of time needed and time spent and research on RT in terms of the effect of attitudinal factors on time use set in the context of predicting students' attainment in the LTL reasoning tasks, we posed four research questions:

*Research Question 1:* How do students' prior abilities as measured by school achievement (GPA) and their motivational attitudes as disclosed in the self-report questionnaire implemented concurrently in the LTL assessment predict their attainment in the LTL reasoning tasks (LTL test score)?

*Research Question 2:* How do the motivational attitudes predict students' TOT in the LTL reasoning tasks when school achievement (GPA) is taken into account?

*Research Question 3:* How is TOT related to students' attainment in the LTL reasoning tasks (LTL test score)?

*Research Question 4:* How does TOT mediate the effect of the motivational attitudes and of school achievement (GPA) on

students' attainment in the LTL reasoning tasks (LTL test score)?

On the basis of the reviewed literature, the following hypotheses were set:

*Hypothesis 1:* School achievement (GPA) is a stronger predictor than motivational attitudes for the LTL test score, but both make independent contributions to it. More specifically, it is expected that higher GPA and stronger mastery attitudes predict a higher LTL test score, whereas detrimental attitudes have a negative effect on it (Klauer, 1988; Spinath et al., 2006).

*Hypothesis 2:* School achievement (GPA) and mastery attitudes are positively related to TOT, whereas detrimental attitudes are negatively related to it (Carroll, 1963; Karweit & Slavin, 1981; Wise & DeMars, 2005).

*Hypothesis 3:* TOT is positively related to the LTL test score (Carroll, 1963; Chang, Plake, & Ferdous, 2005; Goldhammer et al., 2014; Karweit, 1982; Kong et al., 2007).

*Hypothesis 4:* TOT mediates the effect of school achievement (GPA) and motivational attitudes on the LTL test score (Carroll, 1963; Karweit, 1982; Wise & DeMars, 2005; Wise & Kong, 2005).

### Method

#### Participants

The data were drawn from a nationally representative sample of Finnish ninth grade students in spring 2012. Class-based Bernoulli sampling was used with all of the ninth grade classes from each sampled school included in the study. Overall, 8,875 students in 82 schools participated in the assessment, but in the present study, only the data of the 4,249 students (2,153 boys, 2,050 girls, 46 not specified) assigned for CBA are included. The mean age of the students was 15.92 years ( $SD = 0.40$ , range = 14.67–18.57 years).

#### Procedure

The assessment was conducted by a teacher using written instructions. Because of the limited number of computers in schools, the assessment was conducted one class at a time. The students were allocated 90 min for the assessment, a time that had proven sufficient in previous assessments (i.e., Carroll's *allocated time* or *outer constraint*). The number of missing responses in the last task used in this study was only slightly larger than in the first, and only 2% of the students who completed the first task did not complete the last, implying that the time allocated was a close match with the time needed.

#### Measures

**The LTL reasoning tasks.** The six reasoning tasks used in the study, taken from the Finnish LTL test, measure reasoning in different domains. The tasks and their reliabilities are presented in Table 1. The reliabilities were acceptable for all tasks.

Three of the six tasks (Deductive Reasoning, Missing Premises, and Analysis of Relevant and Irrelevant Information) were adapted



Table 1  
*Number of Items and Reliabilities (Cronbach's  $\alpha$ ) of the Cognitive Tasks and the Attitude Scales*

Tasks and scales	Items	$\alpha$
Cognitive tasks		
Deductive Reasoning	6	.60
Missing Premises	10	.61
Relevance of Information	10	.51
Control of Variables	10	.74
Hidden Arithmetical Operators	10	.82
Invented Mathematical Concepts	10	.66
Mastery attitudes		
Agency: Effort	3	.79
Mastery: Extrinsic	3	.89
Importance of School	3	.83
Detrimental attitudes		
Means-Ends: Chance	3	.79
Means-Ends: Ability	3	.67
Self-Handicapping	3	.75

from the Ross Test of Higher Cognitive Processes (Ross & Ross, 1979). In the Deductive Reasoning task, students were presented with six items, each with two premises and followed by three possible conclusions. For each of these, the students had to decide whether they were true or false on the basis of the premises. The student received one point for an item if all three conclusions were answered correctly. The maximum score for the task was 6.

In the Missing Premises task, students were presented with 10 items with one premise and the conclusion given. The students were to choose the second premise from among five alternatives that would make the conclusion valid. Only one of the conclusions was correct. The student received one point for each correct answer. The maximum score for the task was 10.

In the Analysis of Relevant and Irrelevant Information task, which we refer to hereinafter as *Relevance of Information*, the students were presented with 10 arithmetic word problems that contained sufficient, insufficient, or extraneous information for solving the problem. The students were not asked to provide an answer to the arithmetic problems but to assess whether there was just enough, not enough, or even extraneous information given to solve them. The student received one point for each correct answer. The maximum score for the task was 10.

The *Control of Variables* task is a modified version (Hautamäki, 1984) of one of the science reasoning tasks of Shayer (1979), Pendulum, regarding the control of variables. It is based on one of the formal schemata identified by Inhelder and Piaget (1958). The students were presented with 10 items in the form of comparisons set in the world of Formula 1 races with four variables—driver, car, tires, and track—with two alternatives each. The students were to judge whether the single effect of the driver, car, tires, and track could be concluded from the comparison. There were seven comparisons with three or four Yes–No choices for variables and three comparison sets to be complemented. The student received one point if all parts of the item were answered correctly. The maximum score for the task was 10.

The *Hidden Arithmetical Operators* task is based on the quantitative-relational arithmetic operators task of Demetriou and others (Demetriou, Pachaury, Metallidou, & Kazi, 1996; Demetriou, Platsidou, Efklides, Metallidou, & Shayer, 1991). The task

comprised 10 items with one to four operators. For example, “(5 a 3) b 4 = 6. In this task letter a/b stands for: addition (+) / subtraction (–) / multiplication (·) / division (÷)?” The student received one point if all operators in the item were answered correctly. The maximum score for the task was 10.

The *Invented Mathematical Concepts* task is a modified group version of Sternberg's Triarchic Test (H version) Creative Number scale (Sternberg, Castejon, Prieto, Hautamäki, & Grigorenko, 2001), in which arithmetical operators are conditionally defined depending on the value of the digits they combine (e.g., if  $a > b$ , *lag* stands for subtraction, else for multiplication). The task uses two operators with differing definitions and can comprise several operations in the same equation (“What is 4 *lag* 7 *sev* 10 *lag* 3?”). There were 10 items, each with four multiple-choice alternatives for correct solution. The student received one point for each correct answer. The maximum score for the task was 10.

**Motivational scales.** The six motivational scales used in the study are taken from the motivational-affective battery of the LTL test. They were presented to the students concurrently with the LTL reasoning tasks presented above. The chosen scales fall under two constructs: *mastery attitudes* and *detrimental attitudes*. The scales and their reliabilities are presented in Table 1. All affective items were answered on a 7-point Likert-type scale ranging from 1 (*not true at all*) to 7 (*very true*). The reliabilities of all scales were acceptable.

**Mastery attitudes.** The construct comprises scales from three subfields of motivational theory: achievement goal theory (e.g., Elliot & Dweck, 1988; Harackiewicz et al., 2002), agency beliefs (e.g., Chapman, Skinner, & Baltes, 1990), and the internalized value of education (cf. Ryan & Deci, 2000). From achievement goal theory, the construct *mastery extrinsic orientation* (e.g., “Getting good grades at school is important to me.”) was included, tapping into the internalized value of high attainment that Ryan and Deci (2000) called “identification” or “integration” in their taxonomy of human motivation. Of the trio of means–ends beliefs, agency beliefs, and control expectancies, the construct *agency: effort* (e.g., “I work hard to do well at school”) was included, because of its direct reference to learning as an activity. Regarding the internalized value of education, the construct *importance of school* was included, covering students' views on the relevance of school and studying in general (e.g., “I think we learn useful and important things at school”).

**Detrimental attitudes.** Of the attitudes detrimental to learning, two were derived from the background of means–ends beliefs: *means–ends: chance* (e.g., “Failure at school is mainly due to bad luck”) and *means–ends: ability* (e.g., “Poor grades are due to lack of ability”; e.g., Niemivirta, 2002). The third scale included in the detrimental attitude construct was for *self-handicapping* (e.g., “I give up easily if my assignments look too demanding”), also deriving from the broader field of achievement goal theory (e.g., Urdan & Midgley, 2001).

**GPA.** To indicate prior ability, students' GPA was calculated using their self-reported school grades in Finnish, mathematics, English, history, and chemistry in the midterm report card received 4 months before the assessment. The scale of the Finnish school grades runs from 4 (*failed*) to 10 (*excellent*).

**TOT.** The TOT was extracted from the CBA log files for each student for each task. The log file registered the time when the task was opened and the time when it was submitted as finished. Each



task comprised an introduction to the task, one or more presolved example items, and six or more items for the student to solve. As all parts of a task were displayed on the same screen, TOT was available only on a whole task basis. The time was counted in seconds.

## Statistical Method

In studying the effects presented in the research questions and the hypotheses, structural equation modeling (SEM) was used as it can incorporate all the effects into one simultaneous analysis. For structural equation modeling, Mplus 6.0 (Muthén & Muthén, 1998–2010) was used. Because the variables were close to normally distributed (skewness and kurtosis between  $-1$  and  $1$ ) except for the time variables before logarithmic transformations (described below), we used maximum-likelihood estimation. The criterion for an acceptable model fit was comparative fit index (CFI)  $> .900$ , Tucker–Lewis index (TLI)  $> .890$ , and root-mean-square error of approximation (RMSEA)  $< .08$ . For descriptive statistics and outlier definition, we used IBM SPSS Statistics 21.

Hypotheses 1–3 were tested by studying the direct effects in the two specified SEM models: Model 1, in which students' test scores were explained by GPA together with mastery and detrimental motivational attitudes, and Model 2, in which TOT was added to the model to mediate the effects of GPA and the two motivational constructs.

For testing Hypothesis 4, indirect effects in Model 2 were studied, as, according to Zhao, Lynch, and Chen (2010), mediation can be equated with an indirect effect (see also MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). If the direct effect is then not significant, the mediation is full (Zhao et al., 2010). If the direct and the indirect effects are significant and they both are positive or negative the mediation is partial (Zhao et al., 2010). In that case, the direct effect between the independent and dependent variables decreases after the mediator variable is added into the model (MacKinnon, Krull, & Lockwood, 2000). If the direct effect is positive and the indirect effect negative or vice versa, the mediation is competitive, which is also called *suppression*. Contrary to partial mediation, the direct effect between the independent and dependent variables is strengthened after the suppressor variable is added to the model (MacKinnon et al., 2000).

To test the significance of the indirect effects, 95% confidence intervals were produced under inspection. If the interval did not contain zero, the effect was considered significant. Because indirect effects are products of two (or more) effects, their standard errors and consequently their confidence intervals cannot be obtained in a straightforward manner. In Cheung and Lau's (2008) simulation study, the standard errors and confidence intervals for indirect effects were most accurately estimated with a bias-corrected bootstrap method. In the present study, we used this method with 1,000 bootstrap replicates.

## Results

### Descriptives

**TOT.** Because there is little empirical literature regarding the properties and the use of task-based time data in CBA, TOT was

first studied independently from other measures. The distributions are presented in Figure 1.

For determining possible outliers, graphical inspection (see Kong, Wise, & Bhola, 2007; Wise, 2006) was used as the skewed distributions do not allow for methods based on standard deviation (Cousineau & Chartier, 2010) and because outlier values would affect the standard deviations even after normalization of distributions through transformations (Leys, Ley, Klein, Bernard, & Licata, 2013). On basis of the graphical visualization of the distributions, in four of the six tasks, there were outliers whose TOT was considerably longer than that of the other test takers. This might be due to the student not submitting a task before opening a new one in a new window and only later returning to submit the original one. Cutting points at the higher end were defined separately for each task, on the basis of where the relatively even distribution ended: In Deductive Reasoning, three students were categorized as outliers with  $TOT \geq 1,800$  s; in Missing Premises, four students had  $TOT \geq 2,446$  s; in Relevance of Information, six students had  $TOT \geq 1,710$  s; and in Control of Variables, 13 had  $TOT \geq 1,560$  s. In Arithmetical Operations and Mathematical Concepts, no student could be categorized as an outlier, as the TOT was relatively evenly distributed up to 2,972 s and 2,880 s, respectively. Unlike in RT studies where rapid guessing is a concern because of its negative effect on reliability and validity (e.g., Kong et al., 2007; Wise, 2006), no limit for outliers at the lower end of the time distribution was defined. This reflects the focus of the study, with rapid submission seen as one form of a lack of motivational engagement in the assessment tasks.<sup>2</sup>

The outliers' TOT was coded as missing information, which were listwise excluded from the descriptive analyses. The descriptive statistics for each time variable are presented in Table 2.

The distributions of times were skewed, so the measures were transformed (IBM SPSS 21, LG10) into logarithmic scales. This brought the measures to the recommended limits for maximum-likelihood estimation (see Kline, 2005).

**LTL reasoning tasks and the motivational attitude constructs.** The descriptive statistics for all but the time variables are presented in Table 3. The unit of measurement in the reasoning tasks was the percentage of correctly solved items. For the items of the six motivational scales, the raw scores of the 7-point Likert-type scale were used. The scale for the school grades is from 4 (*failed*) to 10 (*excellent*). All variables were almost normally distributed (skewness and kurtosis between  $-1$  and  $1$ ) and the means were close to those in earlier national studies (e.g., Kupiainen, Marjanen, Vainikainen, & Hautamäki, 2011).

### SEM

To test the four hypotheses, two structural equation models were tested on the CBA sample of students. To address Hypothesis 1, in Model 1, attainment in the LTL reasoning tasks (test score) was predicted by prior ability (GPA) and by mastery and detrimental

<sup>2</sup> We ran an auxiliary analysis, filtering out students with very short TOT by the standard deviation method after logarithmic transformation of the time variables as presented by Cousineau and Chartier (2010). The results did not change substantially from the full data, indicating that the inclusion of these students in the continuum of TOT is well founded.



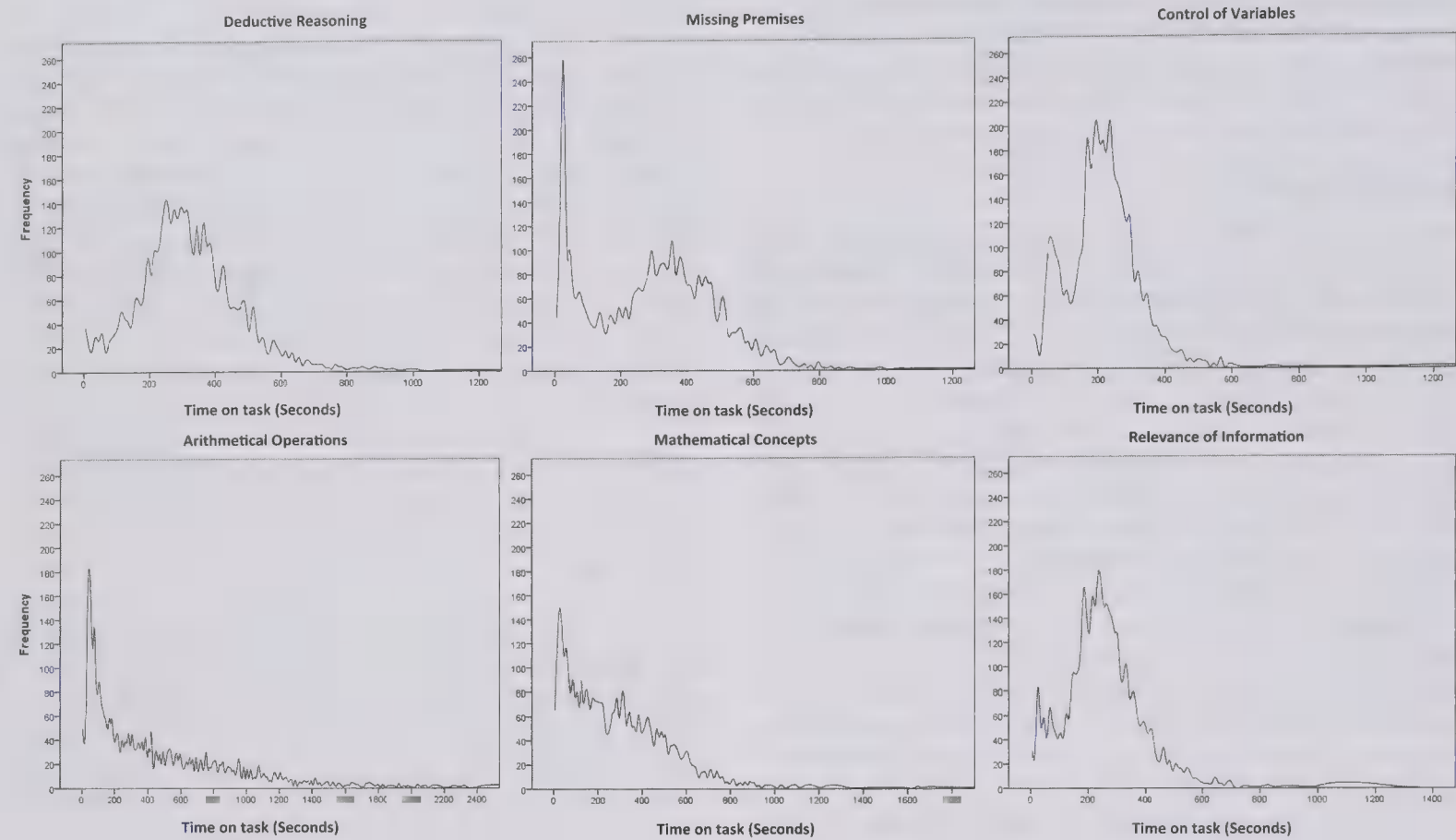


Figure 1. Time on task distributions for the six learning to learn reasoning tasks.

attitudes. To address Hypotheses 2–4 regarding the role of TOT, a TOT construct was added in Model 2 as a mediating factor.

In the two models, the second-order latent factors of mastery and detrimental attitudes were calculated by first regressing the three-item scales on first-order factors of *Mastery: Extrinsic, Agency: Effort*, and *Importance of School* (mastery attitudes) and *Means: Chance, Means: Ability*, and *Self-Handicapping* (detrimental attitudes). The 2 second-order factors were then used to predict the latent factor of the LTL test score, which comprised the scores of the six LTL reasoning tasks. Self-reported school grades in five subjects (Finnish, mathematics, English, history, and chemistry) were regressed on the first-order factor of GPA, also used as a predictor of LTL test score. The fit indices for all of these measurement models were at least acceptable ( $CFI = .957-.989$ ,  $TLI = .920-.967$ ,  $RMSEA = .063-.079$ ) except for the chi-square

statistic (136.718–499.037,  $df = 5-24$ ,  $p < .001$ ), which was expected to be significant because of the large sample size. The latent TOT variable was added to the second model as a mediating factor, comprising the logarithmic TOT variables for each task. To achieve a just acceptable model fit ( $CFI = .965$ ,  $TLI = .895$ ,  $RMSEA = .108$ ) for this measurement model, we had to let two pairs of error terms correlate.

First, Model 1 was fitted to the data (see Figure 2). The fit indices were acceptable ( $CFI = .932$ ,  $TLI = .925$ ,  $RMSEA = .051$ ) except for the chi-square statistic (4,337.273,  $df = 365$ ,  $p < .001$ ). Because of the large sample size and the number of variables in the model, a significant chi-square statistic was expected. Even then, the chi-square value was quite large. For Model 1, the residual correlations were 0.05 on average. Nearly 34% (138/406) of the residuals had an absolute value above 0.05 and 11% exceeded

Table 2  
Descriptive Statistics of the Time on Task Variables Before Logarithmic Transformations

Task	N	Min	Max	M	SD
Deductive Reasoning	4,222	4	1,370	320.65	153.84
Missing Premises	4,120	2	1,658	295.66	197.61
Relevance of Information	4,211	3	1,287	254.22	133.67
Control of Variables	4,138	3	1,321	215.74	110.14
Hidden Arithmetical Operations	4,185	3	2,972	450.08	444.15
Invented Mathematical Concepts	4,206	3	2,880	294.74	258.57

Note. The times were measured in seconds. N = number of students; Min = shortest time on task; Max = longest time on task.



Table 3  
*Descriptive Statistics of the Six Learning-to-Learn (LTL)  
Reasoning Tasks, Three Mastery Attitudes, Three Detrimental  
Attitudes, and Five School Grades in the Models*

Tasks and scales	<i>N</i>	<i>M</i>	<i>SD</i>
LTL reasoning tasks			
Deductive Reasoning	4,243	49.90	27.88
Missing Premises	4,105	41.41	20.94
Control of Variables	4,148	33.47	22.56
Relevance of Information	4,231	36.98	13.81
Hidden Arithmetical Operators	4,200	45.98	24.09
Invented Mathematical Concepts	4,164	25.01	22.97
Mastery attitudes			
Agency: Effort	4,230	4.61	1.19
Mastery: Extrinsic	4,230	5.22	1.23
Importance of School	4,249	4.52	1.18
Detrimental attitudes			
Means–Ends: Chance	4,230	2.39	1.25
Means–Ends: Ability	4,230	3.60	1.15
Self-Handicapping	4,230	3.94	1.27
School grades			
Finnish	4,239	7.78	1.22
Mathematics	4,239	7.52	1.43
English	4,224	7.75	1.32
History	4,234	7.79	1.30
Chemistry	4,238	7.59	1.35

Note. *N* = number of students.

0.10. Finally, 4% of the residuals—mostly for items measuring detrimental attitudes—had an absolute value of 0.15 or larger. The items were, however, kept in the analyses, as they were considered a necessary part in the theoretical framework of the current study.

**Hypothesis 1.** We expected that prior school achievement (GPA) and mastery attitudes would predict the LTL test score positively, whereas detrimental attitudes would have a negative effect on it. Moreover, we expected GPA to be the strongest predictor of the LTL test score. GPA (calculated from self-reported school grades received 4 months prior to the data collection) was indeed the strongest predictor, and the relation between GPA and the LTL test score was strongly positive ( $\beta = .65$ ,  $SE = .02$ ). This result supported earlier literature regarding the role of both prior cognitive ability in explaining school achievement (e.g., Deary, Strand, Smith, & Fernandes, 2007; Rohde & Thompson, 2007) and GPA as an indicator for later cognitive performance (Atkinson & Geiser, 2009; Gustafsson & Carlstedt, 2006; Thorsen & Cliffordson, 2012). Furthermore, there was a positive correlation between GPA and mastery attitudes ( $r = .63$ ,  $SE = .01$ ), whereas detrimental attitudes correlated negatively with GPA ( $r = -.43$ ,  $SE = .02$ ) and with mastery attitudes ( $r = -.30$ ,  $SE = .02$ ). Detrimental attitudes also had a negative relation with the LTL test score ( $\beta = -.31$ ,  $SE = .02$ ). All correlations were statistically significant.

Contrary to Hypothesis 1, the relation of mastery attitudes and the LTL test score was close to zero when students' GPA and detrimental attitudes were controlled for. Despite their positive bivariate correlation ( $r = .43$ ), the path coefficient between mastery attitudes and the LTL test score was negative ( $\beta = -0.08$ ,  $SE = .02$ ).

Together, GPA and mastery and detrimental attitudes explained 61% of the variance in the LTL test score. Thus, even without

taking into account TOT, the model predicted the LTL test score fairly well. Hypothesis 1 was supported regarding the effects of GPA and detrimental attitudes but not for the expected positive effect of mastery attitudes on the test score after controlling for GPA.

To address Hypotheses 2, 3, and 4, we fit Model 2, in which TOT was added to the set of constructs already used in Model 1, to the data (cf. Figure 3). It provided a just acceptable fit ( $CFI = .901$ ,  $TLI = .891$ ,  $RMSEA = .056$ ),  $\chi^2(544) = 7,853.248$ ,  $p < .001$ ). Again, the absolute values of nonredundant residual correlations were .05 on average. About 36% of the residuals were above 0.05 in absolute value and 10% were larger than 0.10. Finally, 4% of the residuals exceeded 0.15 in absolute value. Again, almost all of the large residuals were associated with the items measuring detrimental attitudes and they were nevertheless kept in the analysis for theoretical and validity reasons.

**Hypothesis 2.** The second hypothesis was that higher GPA and stronger mastery attitudes would both make students invest more time in the assessment tasks, whereas a high level of detrimental attitudes would make them invest less time. From Model 2, we can see that detrimental attitudes were a relatively strong negative predictor of TOT ( $\beta = -.41$ ,  $SE = .02$ ), whereas mastery attitudes and GPA predicted it positively ( $\beta = .21$ ,  $SE = .02$ , and  $\beta = .15$ ,  $SE = .02$ , respectively). We conclude that Hypothesis 2 was supported.

**Hypothesis 3.** The third hypothesis was that TOT would be positively related to the LTL test score. Figure 3 shows that adding TOT into the model increased significantly the share of the explained variance in the LTL test score (81% vs. 61% of Model 1) and that the standardized path coefficient from TOT to LTL test score was relatively high ( $\beta = .57$ ,  $SE = .02$ ). Thus, Hypothesis 3 was supported.

**Hypothesis 4.** The fourth hypothesis was that TOT would mediate the effects of motivational attitudes and GPA on the LTL test score. For GPA, the standardized indirect effect on LTL test score was significant but weak ( $\beta = 0.08$ , with a 95% bootstrap confidence interval [0.06, 0.11]). The direct effect was strong even when TOT was taken into account ( $\beta = .65$ ,  $SE = .02$ , of Model 1; cf.  $\beta = .57$ ,  $SE = .02$ , of Model 2). Therefore, the mediating role of TOT between GPA and LTL test score was, at best, partial.

The indirect effect of mastery attitudes on the LTL test score was significant but relatively weak ( $\beta = 0.12$  with a 95% bootstrap confidence interval [0.09, 0.15]). The direct effect was also significant ( $\beta = -.20$ ,  $SE = .02$ ) but negative. This, and the fact that in Model 1 the direct effect was weaker ( $\beta = -.08$ ,  $SE = .02$ ), indicates a suppression effect. In Model 2, the total effect of mastery attitudes on the LTL test score ( $0.12 + -0.20 = -0.08$ ) was equal to the direct effect in Model 1. Thus, adding TOT in the model did not alter the overall influence of mastery attitudes on the LTL test score. Moreover, even if mastery attitudes predicted TOT positively, this increase in effort did not seem to convert into a better achievement in the LTL test.

For detrimental attitudes, the mediating role of TOT was clear: The standardized indirect effect of detrimental attitudes on the LTL test score was  $-0.23$  ( $SE = 0.02$ ) with a 95% bootstrap confidence interval  $[-0.26, -0.20]$ . The direct effect was significant but weak ( $\beta = -.07$ ,  $SE = .02$ ), whereas in Model 1 it was much stronger ( $\beta = -.31$ ,  $SE = .02$ ). Because the direct effect in Model 2 was practically nonexistent, although significant, this can



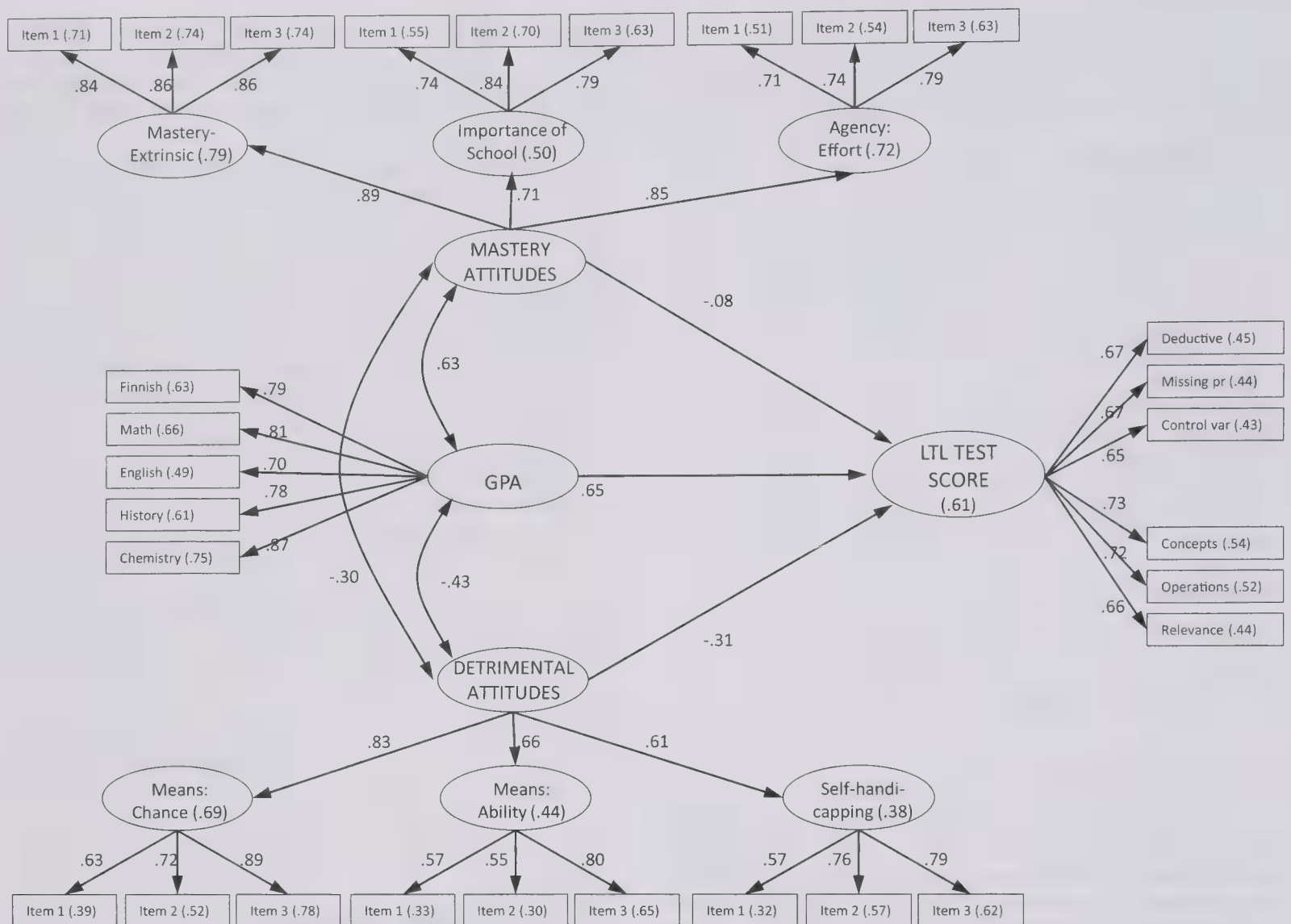


Figure 2. Model 1: Predicting learning-to-learn (LTL) test scores by grade point average (GPA), mastery attitudes, and detrimental attitudes. All of the coefficients are statistically significant,  $p < .001$ . Numbers in parentheses after variable names indicate variance accounted for. Residual variances are not displayed.

(almost) be seen as a case of full mediation. We conclude that Hypothesis 4 was fully supported for detrimental attitudes and partially for GPA and mastery attitudes.

## Discussion

The objective of the study was to investigate the role of TOT in a computer-based low-stakes assessment of cross-curricular LTL reasoning tasks. Theoretically, the study relates to two traditions of time-related learning research: Carroll's (1963) classical model on TOT in learning and the newer assessment-related research on RT (Schnipke & Scrams, 1997). This dual background reflects the double focus of the study on the impact of time use and on factors affecting time use. Accordingly, the findings of the study contribute to both strands of research, and they are of special importance for the growing field of cross-curricular assessment and low-stakes testing.

In this study, we first investigated the role students' prior ability, as indicated by GPA, and their mastery and detrimental motivational attitudes, as disclosed in a self-report questionnaire, had on their attainment in the cross-curricular LTL reasoning tasks. On

the basis of earlier literature, we hypothesized that GPA and mastery attitudes would predict higher attainment (LTL test score), whereas detrimental attitudes would predict a lower LTL test score, and that the effect of GPA would be the strongest. After this, the role of the time students spent on the assessment tasks was explored, using the log data collected in the computer-based assessment. It was hypothesized that higher GPA and stronger mastery attitudes would make students invest more time in the assessment tasks, whereas a high level of detrimental attitudes would make them invest less time. Furthermore, it was expected that TOT would be positively related to the LTL test score and that TOT would mediate the effects of motivational attitudes and GPA on the LTL test score.

To investigate the hypotheses, we specified and fitted two structural equation models to a nationally representative data of 4,249 Finnish ninth grade students. In Model 1, to explore the first research question, the LTL test score was predicted by GPA and by mastery and detrimental motivational attitudes. The hypothesis regarding the role of GPA and of detrimental attitudes was supported, with higher GPA predicting a better LTL test score and



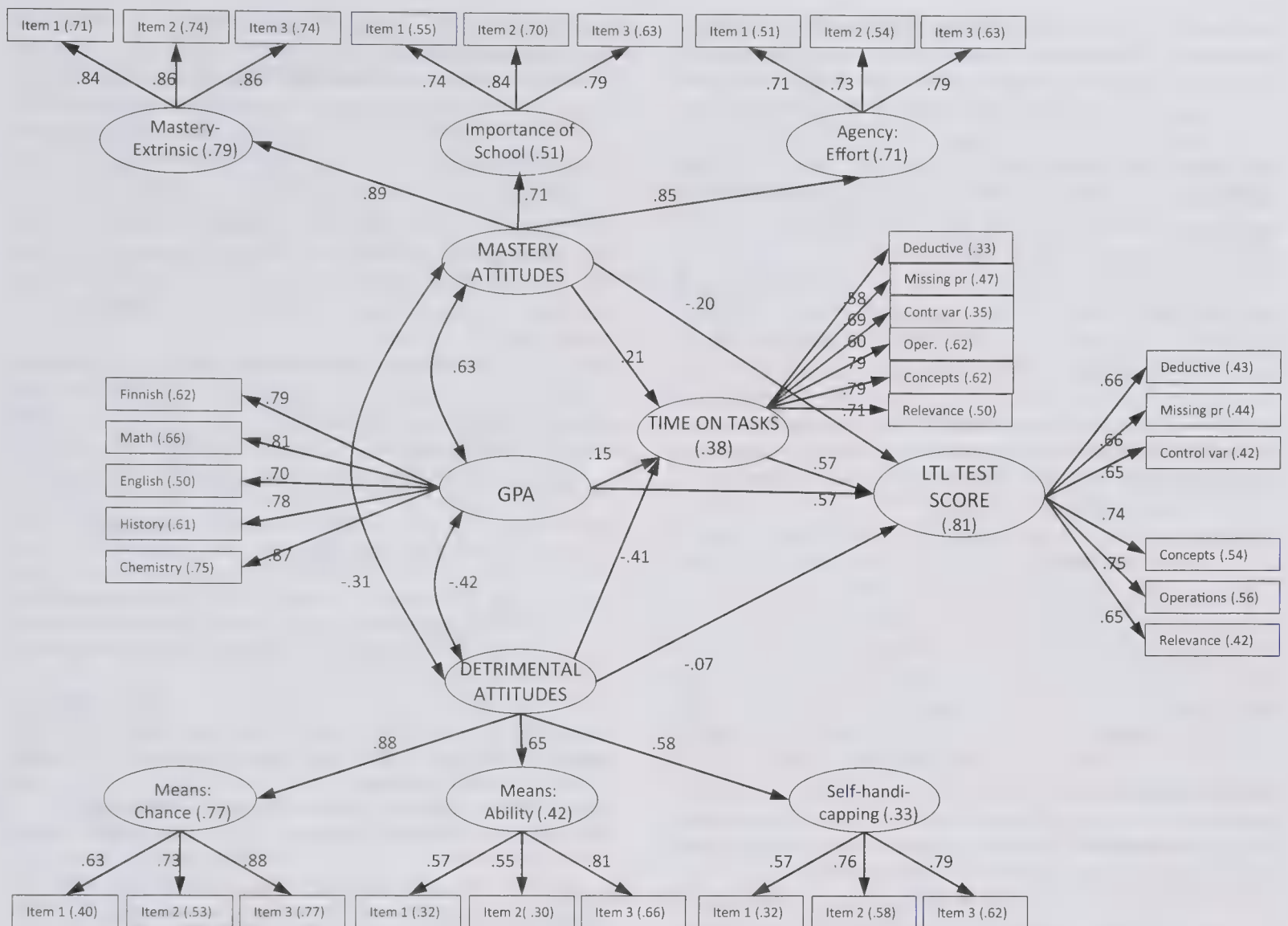


Figure 3. Model 2: Predicting learning-to-learn (LTL) test scores by grade point average (GPA), mastery attitudes, and detrimental attitudes, taking into account the time the students invest in tasks. All of the coefficients are statistically significant,  $p < .001$ . Numbers in parentheses after variable names indicate variance accounted for. Residual variances are not displayed.

detrimental attitudes a weaker LTL test score and the predictive power of GPA being far stronger. Contrary to the hypothesis, however, mastery attitudes had a very weak negative direct effect on the LTL test score despite the positive correlation between the two when GPA was not accounted for. This was interpreted to indicate that students' mastery attitudes get fully rewarded in their GPA, indicating the central role of mastery attitudes in the process where students build their subject-specific achievement through the use of general cognitive ability (cf. Adey et al., 2007).

In Model 2, TOT was added to Model 1 as a mediating factor. First, the relation of GPA and the two attitudinal constructs to TOT was studied. It was confirmed that GPA and mastery attitudes predicted TOT positively but their effect in explaining TOT was weaker than expected. Furthermore, as hypothesized, detrimental attitudes predicted TOT negatively and their effect in explaining TOT was relatively strong.

The comparison of Model 1 and Model 2 confirmed that TOT plays a central role in explaining the LTL test score even when prior school achievement (GPA) is taken into account, increasing the explained variance in the LTL test score from 61% to 81%. It

was also confirmed that TOT mediates the effects of GPA and detrimental attitudes on the LTL test score. For detrimental attitudes, the mediating effect of TOT was almost full, whereas for GPA, it was relatively weak. However, regarding Carroll's notion of the relation of time needed and time used, this was to be expected. The adding of TOT to the model did not change the total effect of mastery attitudes on LTL test score. Moreover, even if mastery attitudes affected TOT positively, this increase in effort did not seem to convert into a better achievement in the LTL test.

Overall, the results support the general finding of the strong positive relation between cognitive ability and school achievement (Deary et al., 2007; Rohde & Thompson, 2007), even if in the present study the direction of prediction was from prior school achievement to attainment in the assessment tasks. The results also support the understanding of the role of motivation and other affective factors on school achievement (Deci & Ryan, 2000; Harackiewicz et al., 2002; Little, Lopez, Oettingen, & Baltes, 2001).

By providing empirical evidence regarding the relations between motivational attitudes, school achievement, and cross-



curricular LTL reasoning skills, the findings make an important contribution to the limited literature on the impact of affective factors on school achievement when controlling for students' general cognitive ability (Gagné & St Père, 2002; Spinath et al., 2006). Moreover, the reversed focus of the study to predict students' attainment in the LTL reasoning tasks instead of school achievement sheds unexpected new light on the mutual relations of motivational attitudes, cognitive ability, and school achievement (cf. Demetriou, Spanoudis, & Mouyi, 2011).

The confirmed relations between GPA, TOT, and students' LTL test score, as well as the added explanative value of Model 2 compared with Model 1 in regard to students' attainment in the LTL reasoning tasks, provide empirical support to Carroll's (1963) model on TOT. The evidence is of special value, as the object of prediction was students' attainment in tasks requiring on-the-spot learning and application of novel rules, and the time spent was measured using the CBA log file, not commonly used earlier in research based on Carroll's construct.

Furthermore, the strong mediating role of TOT in the relation between detrimental attitudes and the LTL test score gives support to Wise and Kong's (2005) interpretation of RT as an indicator for student effort. Yet, although they saw students' use of solution versus rapid guessing behavior to be related to self-reported effort and to explain differences in test scores, they did not look for further explanations for these differences in RT effort. The findings of the present study help to answer this question. Regarding the current widespread use of low-stakes assessment in national and international benchmarking, the finding of the strong negative impact of detrimental attitudes on students' TOT and consequently on their attainment is of prime importance. In this, the study makes an important extension to Wise and his colleagues' (e.g., Wise, 2009; Wise & DeMars, 2005, 2008; Wise & Kong, 2005) research regarding the threat low effort presents to the validity and reliability of assessment data.

The lack of direct effect of mastery attitudes on the LTL test score despite its strong relation with GPA underlines the role of mastery attitudes in aligning students' use of cognitive ability with the goals of the school, rewarded in better grades. This can be seen as just one phase of the continuous cycle of the further development of cognitive ability through engagement in subject-specific learning (cf. Adey et al., 2007; Gustafsson & Carlstedt, 2006). In this, the results support the claim of the Finnish LTL framework of these cross-curricular 21st century skills being fostered through subject-specific teaching and requirements, and they provide one answer to Demetriou et al.'s (2011) call for the education of early adolescents to focus on the development of thinking and problem-solving skills.

The study was built on Carroll's (1963) concept of TOT, but through its context of assessment and the use of CBA log data, it also related to recent research on RT. One way forward on this path of combining the two would be to study students' time investment by separating the time students need to read and assimilate the task instruction from the time they use to solve each item (RT; Chang et al., 2005). Later, this could be used to develop monitoring systems for computer-based learning, allowing teachers to better follow individual students' pace and quality of learning. This would be of special help in supporting struggling students and could be used in the development of adaptive learning and assessment programs. It would also allow the teacher to better

monitor students' TOT behavior and the relation of their motivational attitudes to learning.

The present study confirmed that TOT mediates the effects of detrimental motivational attitudes on test attainment. A next step could be to test the models specified in this study with samples of younger students to see whether TOT would provide a tool for disclosing the effect of their more immature self-awareness on the modeling of the development of the relations of cognitive competence, motivational attitudes, and school achievement (cf. Demetriou & Kazi, 2006; Demetriou et al., 2011; Harter, 1999; for a discussion of response bias, see Bachman & O'Malley, 1984; Buckley, 2009).

The juxtaposition of the two models revealed the role TOT plays in success and its relation to other factors relevant to performance and new learning. The findings point out the advantage CBA log files offer in addressing time investment as a factor crucial for all learning but also for reliable assessment and benchmarking.

## References

- Adey, P., Csapó, B., Demetriou, A., Hautamäki, J., & Shayer, M. (2007). Can we be intelligent about intelligence? Why education needs the concept of plastic general ability. *Educational Research Review*, 2, 75–97. doi:10.1016/j.edurev.2007.05.001
- America Achieves. (2013). *Middle class or middle of the pack? What can we learn when benchmarking U.S. schools against the world's best?* Retrieved from <http://www.americaachieves.org/docs/OECD/Middle-Class-Or-Middle-Of-Pack.pdf>
- Atkinson, R. C., & Geiser, S. (2009). Reflections on a century of college admission tests. *Educational Researcher*, 38, 665–676. doi:10.3102/0013189X09351981
- Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly*, 48, 491–509. doi:10.1086/268845
- Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, 10, 342–363. doi:10.1080/15305058.2010.508569
- Bloom, B. S. (1980). The new direction in educational research: Alterable variables. *The Journal of Negro Education*, 49, 337–349. doi:10.2307/2295092
- Buckley, J. (2009, June). *Cross-national response styles in international educational assessments: Evidence from PISA 2006*. Paper presented at the National Center for Education Statistics Conference on the Program for International Student Assessment (PISA): What We Can Learn From PISA, Washington, DC. <https://edsurveys.rti.org/PISA/>
- Carroll, J. B. (1963). A model of school learning. *Teachers College Record*, 64, 723–733.
- Carroll, J. B. (1989). The Carroll model: A 25-year retrospective and prospective view. *Educational Researcher*, 18, 26–31. doi:10.3102/0013189X018001026
- Cattell, R. B. (1987). *Intelligence: Its structure, growth and action*. Amsterdam, the Netherlands: Elsevier.
- Chang, S.-R., Plake, B. S., & Ferdous, A. A. (2005). *Response times for correct and incorrect item responses on computerized adaptive tests* [Paper presented at the 2005 annual meeting of the American Educational Research Association (AERA), Montréal, Canada]. Retrieved from <http://www.eric.ed.gov/>
- Chapman, M., Skinner, E. A., & Baltes, P. B. (1990). Interpreting correlations between children's perceived control and cognitive performance: Control, agency or means-ends beliefs. *Developmental Psychology*, 26, 246–253. doi:10.1037/0012-1649.26.2.246



- Cheung, G. W., & Lau, R. S. (2008). Testing mediation and suppression effects of latent variables: Bootstrapping with structural equation models. *Organizational Research Methods*, 11, 296–325. doi:10.1177/1094428107300343
- Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: A review. *International Journal of Psychological Research*, 3(1), 58–67.
- Csapó, B. (2007). Research into learning to learn through the assessment of quality and organization of learning outcomes. *The Curriculum Journal*, 18, 195–210. doi:10.1080/09585170701446044
- Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35, 13–21. doi:10.1016/j.intell.2006.02.001
- Deci, E. L., & Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11, 227–268. doi:10.1207/S15327965PLI110401
- Demetriou, A., & Kazi, S. (2006). Self-awareness in *g* (with processing efficiency and reasoning). *Intelligence*, 34, 297–317. doi:10.1016/j.intell.2005.10.002
- Demetriou, A., Pachaury, A., Metallidou, Y., & Kazi, S. (1996). Universals and specificities in the structure and development of quantitative-relational thought: A cross-cultural study in Greece and India. *International Journal of Behavioral Development*, 19, 255–290. doi:10.1080/016502596385785
- Demetriou, A., Platsidou, M., Efklides, A., Metallidou, Y., & Shayer, M. (1991). The development of quantitative-relational abilities from childhood to adolescence: Structure, scaling, and individual differences. *Learning and Instruction*, 1, 19–43.
- Demetriou, A., Spanoudis, G., & Mouyi, A. (2011). Educating the developing mind: Towards an overarching paradigm. *Educational Psychology Review*, 23, 601–663. doi:10.1007/s10648-011-9178-3
- Elliott, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology*, 54, 5–12. doi:10.1037/0022-3514.54.1.5
- Gagné, F., & St Père, F. (2002). When IQ is controlled, does motivation still predict achievement? *Intelligence*, 30, 71–100. doi:10.1016/S0160-2896(01)00068-X
- Gettlinger, M. (1985). Time allocated and time spent relative to time needed for learning as determinants of achievement. *Journal of Educational Psychology*, 77, 3–11. doi:10.1037/0022-0663.77.1.3
- Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, 106, 608–626. doi:10.1037/a0034716
- Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational settings—Something beyond *g*: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology*, 105, 364–379. doi:10.1037/a0031856
- Grek, S. (2009). Governing by numbers: The PISA ‘effect’ in Europe. *Journal of Education Policy*, 24, 23–37. doi:10.1080/02680930802412669
- Gustafsson, J.-E., & Carlstedt, B. (2006, August). *Abilities and grades as predictors of achievement: The encapsulation theory*. Paper presented at the 114th Annual Convention of the American Psychological Association, New Orleans, LA.
- Harackiewicz, J. M., Barron, K. E., Pintrich, P. R., Elliot, A. J., & Thrash, T. M. (2002). Revision of achievement goal theory: Necessary and illuminating. *Journal of Educational Psychology*, 94, 638–645. doi:10.1037/0022-0663.94.3.638
- Harter, S. (1999). *The construction of the self: A developmental perspective*. New York, NY: Guilford Press.
- Hattie, J. (2005). The paradox of reducing class size and improving learning outcomes. *International Journal of Educational Research*, 43, 387–425. doi:10.1016/j.ijer.2006.07.002
- Hautamäki, J. (1984). *Peruskoululaisten loogisen ajattelun mittaamisesta ja esiintymisestä* [The measurement and distribution of Piagetian stages of thinking in Finnish comprehensive school]. Joensuu, Finland: University of Joensuu, Department of Social Sciences.
- Hautamäki, J., Arinen, P., Eronen, S., Hautamäki, A., Kupiainen, S., Lindblom, B., . . . Scheinin, P. (2002). *Assessing learning-to-learn: A framework*. Retrieved from the Finland National Board of Education website: [http://www.oph.fi/download/47716\\_learning.pdf](http://www.oph.fi/download/47716_learning.pdf)
- Hautamäki, J., Kupiainen, S., Arinen, P., Hautamäki, A., Niemivirta, M., Rantanen, P., & Scheinin, P. (2006). Learning-to-learn assessment in Finland: Versatile tools to monitor and improve effectiveness and equity of the education system. In R. Jakku-Sihvonen & H. Niemi (Eds.), *Research-based teacher education in Finland: Reflections by Finnish teacher educators* (pp. 189–202). Jyväskylä, Finland: Finnish Educational Research Association.
- Hoskins, B., & Fredriksson, U. (2008). *Learning to learn: What is it and can it be measured?* (JRC Scientific and Technical Report EUR 23432 EN). Ispra, Italy: European Commission, Joint Research Centre, Centre for Research on Lifelong Learning.
- Inhelder, B., & Piaget, J. (1958). *The early growth of logic in the child*. London, United Kingdom: Routledge & Kegan Paul.
- Karweit, N. (1982). *Time on task: A research review* [Report No. 332]. Retrieved from <http://www.eric.ed.gov/PDFS/ED228236.pdf>
- Karweit, N., & Slavin, R. E. (1981). Measurement and modeling choices in studies of time and learning. *American Educational Research Journal*, 18, 157–171. doi:10.3102/00028312018002157
- Klauser, K. J. (1988). Teaching for learning-to-learn: A critical appraisal with some proposals. *Instructional Science*, 17, 351–367. doi:10.1007/BF00056221
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.
- Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, 67, 606–619. doi:10.1177/0013164406294779
- Kupiainen, S., Marjanen, J., Vainikainen, M.-P., & Hautamäki, J. (2011). *Oppimaan oppiminen Vantaan peruskouluissa: Kolmas-, kuudes- ja yhdeksäsluokkalaisten oppijoina keväällä 2010* [Learning to learn in comprehensive schools in Vantaa: 3rd, 6th and 9th graders as learners in spring 2010]. Department of Education, University of Helsinki, Vantaa, Finland.
- Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, 53, 359–379.
- Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49, 764–766. doi:10.1016/j.jesp.2013.03.013
- Little, T. D., Lopez, D. F., Oettingen, G., & Baltes, P. B. (2001). A comparative-longitudinal study of action-control beliefs and school performance: On the role of context. *International Journal of Behavioral Development*, 25, 237–245. doi:10.1080/01650250042000258
- MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, 1, 173–181. doi:10.1023/A:1026595011371
- MacKinnon, D. P., Lockwood, C. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7, 83–104. doi:10.1037/1082-989X.7.1.83
- Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.
- Niemivirta, M. (2002). Individual differences and developmental trends in motivation: Integrating person-centred and variable-centred methods. In



- P. R. Pintrich & M. L. Maehr (Eds.), *New directions in measures and methods: Advances in motivation and achievement* (Vol. 12, pp. 241–275). Amsterdam, the Netherlands: JAI Press.
- Olson, D. (2003). *Psychological theory and educational reform: How school remakes mind and society*. Cambridge, United Kingdom: Cambridge University Press.
- Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence*, 35, 83–92. doi:10.1016/j.intell.2006.05.004
- Ross, J. D., & Ross, C. M. (1979). *Ross test of higher cognitive processes*. Novato, CA: Academic Therapy.
- Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25, 54–67. doi:10.1006/ceps.1999.1020
- Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford, United Kingdom: Pergamon Press.
- Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, 34, 213–232. doi:10.1111/j.1745-3984.1997.tb00516.x
- Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Mahwah, NJ: Erlbaum.
- Shayer, M. (1979). Has Piaget's construct of formal operational thinking any utility? *British Journal of Educational Psychology*, 49, 265–276. doi:10.1111/j.2044-8279.1979.tb02425.x
- Snow, R. E. (1990). New approaches to cognitive and conative assessment in education. *International Journal of Educational Research*, 14, 455–473.
- Spinath, B., Spinath, F. M., Harlaar, N., & Plomin, R. (2006). Predicting school achievement from general cognitive ability self-perceived ability and intrinsic value. *Intelligence*, 34, 363–374. doi:10.1016/j.intell.2005.11.004
- Sternberg, R., Castejon, J. L., Prieto, M. D., Hautamäki, J., & Grigorenko, E. (2001). Confirmatory factor analysis of the Sternberg Triarchic Abilities Test in three international samples. *European Journal of Psychological Assessment*, 17, 1–16. doi:10.1027//1015-5759.17.1.1
- Sundre, D. L., & Wise, S. L. (2003, April). 'Motivation filtering': An exploration of the impact of low examinee motivation on the psychometric quality of tests. Paper presented at the National Council on Measurement in Education Annual Conference, Chicago, IL.
- Thorsen, C., & Cliffordson, C. (2012). Teachers' grade assignment and the predictive validity of criterion-referenced grades. *Educational Research and Evaluation: An International Journal on Theory and Practice*, 18, 153–172. doi:10.1080/13803611.2012.659929
- Urdan, T., & Midgley, M. (2001). Academic self-handicapping: What we know and what more there is to learn. *Educational Psychology Review*, 13, 115–138. doi:10.1023/A:1009061303214
- Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, 68, 5–24. doi:10.1177/0013164407305592
- Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, 19, 95–114. doi:10.1207/s15324818ame1902\_2
- Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *Journal of General Education*, 58, 152–166. doi:10.1353/jge.0.0042
- Wise, S. L., & DeMars, C. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10, 1–17. doi:10.1207/s15326977ea1001\_1
- Wise, S. L., & DeMars, C. (2008, March). *Examinee non-effort and the validity of program assessment results*. Paper presented at the 2008 annual meeting of the National Council on Measurement in Education, New York, NY.
- Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18, 163–183. doi:10.1207/s15324818ame1802\_2
- Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, 37, 197–206. doi:10.1086/651257

Received March 15, 2013

Revision received December 2, 2013

Accepted December 3, 2013 ■



# Computer-Based Assessment of School Readiness and Early Reasoning

Benő Csapó, Gyöngyvér Molnár, and József Nagy  
University of Szeged

This study explores the potential of using online tests for the assessment of school readiness and for monitoring early reasoning. Four tests of a face-to-face-administered school readiness test battery (speech sound discrimination, relational reasoning, counting and basic numeracy, and deductive reasoning) and a paper-and-pencil inductive reasoning test were transferred to an online platform and administered at the beginning of school to samples of first-grade children (the sample sizes were between 364 and 435). Results of the original and the computerized tests were analyzed to explore (a) whether the new scales were identical to the original ones; (b) how the change of media influenced the reliability of the tests; and (c) whether the migration into a new medium affected gender differences. Analyses indicated that measurement invariance held in a strict sense in the case of the inductive reasoning test (the migration did not change the general look of the test or the item types) and only partially for the speech sound discrimination test (neither the item type nor the scoring principle was changed). Measurement invariance did not hold for the 3 remaining tests. In 3 tests—speech sound discrimination, relational reasoning, and deductive reasoning—the online versions demonstrated improved reliability. Only certain items of the numeracy test could be assessed on computer, and the reliability of the shortened test decreased. No differences were found between the 2 versions of the inductive reasoning test. Gender differences were explored for the speech sound discrimination test, and latent analyses indicated that measurement invariance did not hold. Girls' performance was somewhat better, similarly to former face-to-face assessments, where girls performed slightly better than boys. These results encourage further research on the extension of computer-based assessment to early childhood education.

**Keywords:** computer-based assessment, online testing, school readiness, inductive reasoning, early childhood assessment

A large number of studies have highlighted the importance of smooth preschool-to-school transition and the successful first years of schooling from different perspectives. Research has paid increasing attention to identifying the conditions of a successful start in schooling. Among these efforts, creating instruments for assessing school readiness and monitoring development at the beginning of schooling play an important role. A broad range of instruments, including observation protocols, tests, and test batteries, are available, which can be used to assess different aspects of general cognitive development as well as specific precursors of skills learners are expected to master at school. However, many instruments that have been proven valid and reliable under research or pilot conditions turn out to be too complicated to use regularly in schools. Sometimes they are not sufficiently precise if not used under standardized conditions or if not administered by specially trained teachers. In many cases, the time and human resources required to administer and score the tests prevent their frequent use. Technology-based assessment may solve these problems, but administering computerized tests to young chil-

dren before or at the initial stage of formal schooling may raise a number of questions concerning the validity of results obtained through technology-based assessment in young children.

In this article, we explore the possibilities of online testing at the beginning of formal education by comparing traditional and digitized versions of five tests. Four of them are tests from the DIFER (Diagnosztikus FEjlődésvizsgáló Rendszer—diagnostic system for assessing development) school readiness test battery, an instrument with a long developmental history (Nagy, 1980, 1987; Nagy, Józsa, Vidákovich, & Fazekasné Fenyvesi 2004a, 2004b). The fifth is an inductive reasoning test prepared to measure learners' general mental ability. These five instruments measure different psychological attributes, and their computerized versions require different technological solutions. This variety of instruments offers a number of possibilities to analyze the prospects for and limitations of technology-based assessment around the time of the kindergarten-school transition. As we focus on the applicability of technology, we deal only in brief with the general functions of school readiness tests and other instruments used for monitoring children's development during the first school years.

---

This article was published Online First February 17, 2014.

Benő Csapó, MTA-SZTE Research Group on the Development of Competencies, University of Szeged, Szeged, Hungary; Gyöngyvér Molnár and József Nagy, Institute of Education, University of Szeged, Szeged, Hungary.

Correspondence concerning this article should be addressed to Benő Csapó, MTA-SZTE Research Group on the Development of Competencies, University of Szeged, Petőfi sgt. 30-34, Szeged, Hungary. E-mail: csapo@edpsy.u-szeged.hu

## Assessment of School Readiness and Early Development

Mastery of basic literacy and numeracy skills is the main goal of the first school years; therefore, school readiness tests are often composed of tasks that measure precursors to speaking skills, vocabulary, early reading, writing, counting, computing, reasoning (comprehending relations and inferential processes), and the elements of behavior and social skills (attention, following instruc-



tions, and collaborating) that are necessary for working in classroom settings (Konold & Pianta, 2005). Longitudinal research indicates that early (preschool as well as first grade) mathematical and reading skills represent strong predictors of later achievement (Duncan et al., 2007; Hair, Halle, Terry-Humen, Lavelle, & Calkins, 2006; Magnuson, Ruhm, & Waldfogel, 2007; Merrell & Tymms, 2010). For example, Tymms, Jones, Albone, and Henderson (2009) reported correlations ranged from .65 to .80 between kindergarten assessments and mathematics and reading achievements measured in the first and fifth grades.

A number of instruments have been developed to monitor early development (see C. E. Snow & Van Hemel, 2008), but only a few of them are used in regular educational practice due to theoretical and practical constraints. Among the theoretical problems often cited are the difficulties of defining the concept of school readiness and properly determining the purpose of testing. For instance, readiness may mean either readiness to learn or readiness to perform in a school setting (Carlton & Winsler, 1999). If readiness testing is focused on predicting school performance, then it will be the slow developers or low-performing children who are most in need of the developmental influences provided by school that are prevented from entering it (Shepard, 1997). To overcome these difficulties, a more complex conception of school readiness is proposed, a concept that also takes into account children's cognitive, emotional, and social development (Blair, 2002). These issues are less crucial if school readiness tests are used as diagnostic tools and identification of deficiencies is followed by treatment.

Early tests must take into account that the children assessed may not be able to read. Thus, these tests are usually individually administered with stimuli presented and instructions read by test administrators, who also record the answers. This limits the standardization of testing conditions and leaves the process open to subjective influences and interpretations of test takers' responses. Research on the quality of school readiness testing indicates that assessments made by teachers are often biased (as they are less strict with the children) when their conclusions are compared with results from objective assessment instruments (Mashburn & Henry, 2004). Despite these constraints, a number of school readiness assessments are based on the direct observation of children (e.g., the Early Development Instrument; see Guhn, Janus, & Hertzman, 2007). The Performance Indicators in Primary Schools (PIPS) tests are used to monitor children's development in the early years of primary school. Its Baseline Assessment (PIPS BLA) measures early reading; mathematics; phonological awareness; and personal, social, and emotional development on a 5-point scale (Merrell & Bailey, 2012). This instrument was used in a large-scale international study (iPIPS) to compare children's early development in English-speaking countries.

Although the majority of studies on school readiness assessment have focused on the cognitive domain, recent research identified several further factors, which play a crucial role in kindergarten-school transition and later development, such as self-concept, peer status, classroom contexts, and parenting (Bossaert, Doumen, Buyse, & Verschueren, 2011; McWayne, Cheung, Wright, & Hahs-Vaughn, 2012). Although there are still a number of open questions related to certain details of the content of school readiness assessment and the ways their data may be used there is a consensus that the availability of appropriate and easy-to-use measurement instruments is crucial to helping children to begin school

successfully and to identify those who are in need of additional support (K. L. Snow, 2006).

### Computer-Based Assessments and the Context of the Present Study

In educational practice, there may be different forces and interests driving the search for better solutions and applications of technology to replace traditional (face-to-face and paper-and-pencil) forms of assessment. The main factors motivating the use of technology are improving the assessment of already established assessment domains (Csapó, Ainley, Bennett, Latour, & Law, 2012) and measuring constructs that would be impossible or difficult to measure without the means of technology (e.g., Complex Problem Solving; see Greiff, Wüstenberg, & Funke, 2012; Greiff, Wüstenberg, Holt, Goldhammer, & Funke, 2013; Greiff et al., 2013).

Computer-based (CB) and paper-and-pencil (PP) test comparability studies were among the most extensively researched questions over the last two decades. Because of several advantages, CB assessment delivery has been gradually replacing traditional PP delivery as it permits the tailoring of tests to the individual characteristics of learners (e.g., adaptive testing), automated scoring (including promising developments in children's speech recognition) and immediate feedback, the inclusion of innovative item formats (e.g., multimedia elements, simulation, and dynamic items), precise control over the presentation of test stimuli, and reduced costs of test administration (see Price et al., 2009). One of the regular large-scale assessment programs, the Programme for International Student Assessment (PISA), is also gradually shifting from PP to CB assessments. In PISA (2006, 2009, 2012; see Organisation for Economic Co-Operation and Development, 2010, 2011, 2014), CB assessments were offered as international options or took place in one of the innovative domains; in 2015, the major domains (reading, mathematics, and science) will be assessed with computerized instruments.

A number of studies have been conducted in different knowledge and competence domains using a variety of educational tests to examine whether test delivery mode affects children's performance (Clariana & Wallace, 2002; Kingston, 2009; Wang, Jiao, Young, Brooks, & Olson, 2008). The differences between PP and CB test performance in terms of validity and reliability, advantages and disadvantages, and the effects of background variables (gender, race/ethnicity, and technology-related factors, such as computer familiarity; Csapó, Molnár, & Tóth, 2009; Gallagher, Bridgeman, & Cahalan, 2000) have been widely studied and well documented. Most of the recent media effect studies have indicated that PP and CB testing are comparable and that students prefer CB tests to traditional PP testing. Although the research results are inconsistent to some extent, comparability problems are likely to decrease over time as computers become more broadly accessible at schools (Way, Davis, & Fitzpatrick, 2006). Even though there is a lively debate over comparative studies, less attention has been paid to the effects of different delivery modes on different subgroups of the samples, and only a few studies have focused on testing very young learners in a technology-based environment (Carson, Gillon, & Boustead, 2011; Choi & Tinkler, 2002).



The most widely studied subgroup differences are those between girls and boys. Gender differences are routinely analyzed in large-scale assessment programs and are especially relevant in CB testing. The results of previous studies have revealed that the new media slightly changed the pattern of differences compared with the traditional PP assessments. In large-scale international PP assessments, the overall pattern is that boys and girls perform alike or boys do slightly better than girls in mathematics and science, whereas girls perform better than boys in reading. Boys usually perform better on the information-communication technology (ICT) literacy and computer familiarity test, and their better ICT skills may affect the results of CB tests. For instance, in the PISA (2006) study, science was tested in three countries using computers as well, and gender differences varied across the three participating countries on the Computer-Based Assessment of Science. However, boys outperformed girls on average (Organisation for Economic Co-Operation and Development [OECD], 2010). In the PISA (2009) survey, electronic reading was an innovative assessment domain. On the Electronic Reading Assessment, girls outperformed boys on average, and this pattern was the same across all OECD countries (OECD, 2011). Horne (2007) reported similar results in reading and spelling tests. In general, no gender differences were found on the computerized versions, whereas girls outperformed boys on the paper versions of the tests. A study compared the achievement of fifth-grade (11-year-old) primary-school children in inductive reasoning measured by PP and CB in a larger representative sample in Hungary and indicated no achievement differences between boys and girls in PP or CB test results (Csapó et al., 2009). In the context of early testing, analyzing gender differences may be essential for making existing instruments equally usable for boys and girls.

One of the main tasks of current developmental efforts is to migrate well-established face-to-face or PP tests to the new technology. However, whereas a switch to CB delivery is accompanied by some obvious improvements in efficiency, cost-effectiveness, and precision, further research is required to determine potential changes in reliability, ecological validity, applicability, and possible biases when migrating testing to the new medium. Most previous mode effect studies compared PP and CB delivery modes only. In the present study, we explore the differences between individual face-to-face and online testing as well.

### **The Development of the DIFER Test Battery and the Inductive Reasoning Test**

The development of the school readiness test battery, which is at the center of this study, started back in the 1970s, when the first large-scale assessment of young learners in Hungary explored a group of skills necessary for a successful start of schooling. The results of this work (Nagy, 1980) formed the foundations for developing an extensive instrument, the PREFER test battery, administered face to face (FF) by teachers and covering the most essential competencies needed to begin school successfully (Nagy, 1987). After using it in educational practice for more than a decade, it was revised and renewed as the DIFER test battery (Nagy et al., 2004b). It has been used in several large-scale assessments to establish its reliability and predictive validity (Nagy et al., 2004a). Five DIFER tests were used in a longitudinal program, where they were the first instruments administered to a

sample followed for 10 years (Csapó, 2007; Józsa, 2004). Strong correlations were found between the DIFER test and later school achievement. For example, the results of the counting and basic numeracy DIFER test correlated at .60 with a counting test administered at the end of second grade, and the correlation remained .49 (with a mathematical reasoning test) at the end of fifth grade and .48 at the end of eighth grade, with the mathematics test administered within the framework of the National Assessment of Basic Competencies (Csapó, 2013).

In educational practice, the DIFER can be used as a diagnostic instrument. Children are assessed regularly over time, and a record of their development is kept in a booklet. The development of those who lag behind may be stimulated by special purpose exercises. The DIFER is designed so that its administration does not require specific expertise; it can be administered by kindergarten and primary-school teachers. A major drawback of the test battery is that it must be administered face to face and individually. This is especially problematic for primary schools, where it is difficult to fit testing sessions into teachers' and learners' schedules. Another issue is the objectivity of the test administration as teachers may read the instructions to children in slightly different ways, and the scoring of the responses may also vary. An online delivery of prerecorded voice instructions (with texts read by trained speakers) and automated scoring may solve these problems. As two out of the seven DIFER tests (social skills and writing) cannot be immediately digitized, the remaining five (speech sound discrimination, relational reasoning, deductive reasoning, inferential reasoning, and counting skills) were transferred to an online platform in this study. The characteristics of inferential reasoning do not differ much from those of deductive reasoning; thus, we omitted inferential reasoning from the analyses presented in this article.

To increase the variety of the instruments, a PP inductive reasoning test and its digitized version were added to the four remaining DIFER instruments (speech sound discrimination, relational reasoning, counting and basic numeracy, deductive reasoning). The development of the inductive reasoning tests began in the early 1990s (Csapó, 1997). Several PP inductive reasoning tests have been in regular use for almost two decades both for mapping the development of inductive reasoning itself (Csapó, 2007) and for measuring inductive reasoning performance as an indicator of the developmental level of higher order thinking skills (Csapó & Nikolov, 2009). Later on, a second inductive reasoning test was constructed for the early grades, based on Klauer's model of inductive reasoning (Klauer, 1989), and has been used in experiments to assess the effect of training (Molnár, 2011). Using item response theory (IRT) analyses, both tests were equated so that their results were represented on the same scale (common-person methods were used; Molnár & Csapó, 2011). Finally, computerized versions were created for both tests, and the effects of delivery mode were studied by comparing the PP and CB versions (Csapó et al., 2009).

### **Research Questions**

In this study, we explore the possibilities of the application of online assessment in regular educational practice at the beginning of schooling. For this purpose, we apply FF and PP tests that already have established psychometric characteristics; we transfer



them to the new media and compare them by answering three research questions.

1. Does the medium of delivery influence the results, or can the results of the tests be represented on the same scale (i.e., testing of measurement invariance)?

2. If the tests differ between the two media, what influences these differences (i.e., psychometric properties)?

3. Does changing the mode of administration affect gender differences (i.e., latent mean differences between boys and girls)?

### Method

A number of constraints have to be taken into account when carrying out comparative assessments with children entering school using an emerging technology.

1. Although online technology has had a relatively short developmental history, it has been extensively piloted with schoolchildren of different age groups, but not yet with preschool children.

2. To ensure comparability and to prevent the impact of schooling, a very short period is available for testing; schoolchildren may only be assessed at the very beginning of the first school year. (All assessments reported here took place during the first weeks of the first school year.)

3. As ecological validity is a main concern of the research, all assessment occurred in real school settings using the available infrastructure.

These conditions were equally taken into account when the study was designed, data sources selected, and procedures planned.

### Participants

Data for two DIFER tests (relational reasoning test and counting and basic numeracy test) were drawn from an assessment in which all Hungarian children of school-entering age were assessed with the DIFER tests. A subsample was randomly selected for further detailed analyses; we used these data in this study. Two further DIFER tests (speech sound discrimination and deductive reasoning) were administered to different samples representatively drawn from the school-entering population. The PP inductive reasoning test was administered to a further representative sample. In each case, school classes formed the units of selection. The sample sizes and the attributes of the tests administered to the samples are summarized in Table 1.

The digitized versions of all five tests were administered to different samples due to organizational issues. These samples were randomly drawn from first-grade children in Hungarian primary

schools. The online version of the speech sound discrimination test was administered to the same sample as its FF version. In this case, the order of modes was randomized, and there was a 2-week interval between the two testing sessions.

### Instruments

The study is based on five tests that measure different skills essential for later learning. These key skills include (a) speech sound discrimination, a prerequisite of successful reading; (b) the ability to understand the meaning of words that denote relations; (c) number concept and basic counting skills; and basic (d) deductive and (e) inductive reasoning skills, all of which are prerequisites to learning to read and to studying mathematics and science.

An FF or PP version existed for each test, as described in the theoretical part of the present article. For the present study, we constructed electronic versions of the existing instruments, basically by migrating the items to the new platform. The viability and success of the migration depended on the content of the assessment and the item type. In the process of test digitization, one of the central aims was to preserve as many features of the items as possible in order to make the two delivery modes comparable. The paper and screen layouts were identical or as similar as possible.

Two out of the seven DIFER tests could not be implemented in the new medium. The FF social skills test is based on an observation of the children's behavior. This test proved to have high predictive validity in a longitudinal study, but it could not be realized on a computer. The writing test examined fine hand movement (fine motor skills), which is a precondition of learning handwriting. It was not possible to implement this with the available technology. The other FF DIFER tests were converted into CB formats, although some items had to be omitted and the open-ended items were reformulated and converted into multiple-choice items to allow automated scoring. Only items implemented in both media were used in the comparative analyses.

**Speech sound discrimination test.** This test includes 60 items that measure the perception of phonemic contrasts. The test reveals whether children have good hearing and are able to differentiate some critical pairs of phonemes, for instance /v/ - /f/ and /b/ - /p/ (e.g., in the pairs of Hungarian words *vonat-fonat* and *bont-pont*).

In the first part of the original FF version of the test, the administrator read two sentences; each one contained one of the words from the pairs and showed the matching picture depicting the object referred to in the sentence. Then the administrator read only one of the words. The children indicated their answer by pointing to the picture that matched the word the administrator had

Table 1  
*The Samples in the Study and the Attributes of the Tests (Number of Items, Reliability)*

Test	Sample sizes		Number of items	Cronbach's $\alpha$	
	N (FF or PP)	N (CB)		FF or PP	CB
Speech sound discrimination <sup>a</sup>	(FF) 364	364	60	.887	.938
Relational reasoning	(FF) 1,892	426	24	.796	.844
Counting and basic numeracy	(FF) 1,895	435	13	.812	.770
Deductive reasoning	(FF) 424	402	32	.743	.831
Inductive reasoning	(PP) 952	377	37	.855	.856

Note. FF = face-to-face; PP = paper and pencil; CB = computer-based.

<sup>a</sup> The FF and CB tests were administered to the same sample.



read. Finally, the test administrator scored and logged the answers on a scoring sheet. In CB mode, the same pictures were presented on the screen, and instructions were given online by a prerecorded voice. Children used headsets and heard the same voice of a trained speaker. They had to indicate their answer by using the mouse and clicking on the correct picture. An analogous English-language example could be the following: "This is a sheep (showing a picture of a sheep). This is a ship (showing a picture of a ship). Now I will only say one word. Point/click at the picture, which depicts it."

The second part of the test focused on children's phoneme perception in fluent speech and on the correct pronunciation of a word depicted by a picture. Finally, the third and fourth subtests contained pairs of real or pseudowords or words that differed in one phoneme. Test takers had to decide whether the two words in each pair matched or not.

**Relational reasoning test.** Understanding words that denote relations between different objects, attributes, or processes is a precondition of school learning. The DIFER contains four equivalent versions of relational reasoning tests both in FF and in CB mode, each containing 24 items. As their structure was identical, we use only one version in this analysis. In each test, there were eight relation words tied to space (e.g., *inside*, *between*), four relation words encoding quantity (e.g., *odd*, *few*), four relation words referring to actions (e.g., *step on*, *step in*), four relation words related to time (e.g., *earlier*, *later*), and, finally, four different relational expressions encoding physical measures (e.g., "the shortest," "the same length"). Figure 1 shows a sample item from the CB test; the picture was the same in FF mode as well.

With FF administration, the instructions were given and the test administrator scored the answers. Children had to supply their answers by pointing to the matching picture(s). In the CB environment, instructions were given online; students had to provide their answers by using the mouse and clicking on the matching picture(s).

**Counting and basic numeracy skills test.** The original test constructed for FF administration consisted of items that measure the understanding of the meaning of numbers, number relations,

and basic mathematical thinking. Some items were based on oral counting, and, as the online platform is not yet able to handle oral responses, a number of items on the original numeracy test were omitted from the CB version. Only items that test recognition of quantities, numbers, and representations of numbers were kept. Figure 2 illustrates items on the CB version of the test.

The FF and CB data collection proceeded the same way as in the sound discrimination test described before.

**Deductive reasoning test.** Deductive reasoning was measured with 32 open-ended, contextually embedded tasks in FF mode and with 32 multiple-choice tasks in CB mode to make automated scoring practicable and to allow immediate feedback after testing in the latter case. Each task began with two premises (statements), and children had to reach and formulate a logical conclusion. The context of the situations presented to them may have been familiar from everyday life, so it would have been possible for them to use real-world knowledge to formulate their conclusions.

**Inductive reasoning test.** The structure of the inductive reasoning test was based on Klauer's (1989) definition of inductive reasoning. Klauer defined inductive reasoning as discovering regularities by detecting similarities, dissimilarities, or a combination of both, with respect to attributes or relations to or between objects. This involved six classes in total (generalization, discrimination, cross-classification, recognizing relations, discriminating relations, and system formation). The test consisted of 37 figural, nonverbal items belonging to the six subclasses of inductive reasoning described above. Figure 3 illustrates the items on the inductive reasoning test; the same pictures were used both in PP and CB modes.

During the digitization of the test, all features of the items were preserved to make the two versions as similar as possible. For example, in the PP multiple-choice items, children had to circle or underline the letter or the picture, whereas in the CB format, they had to click on the same letter or picture to indicate their answer (see Figure 3).

## Procedure

Two traditional delivery methods, FF and PP testing modes, were used. During FF administration, children were tested individually. The instructions for the items were read by test administrators, most of whom were the children's homeroom teachers, and children's answers were recorded on a scoring sheet. The PP version of the inductive reasoning test was taken in the children's regular classroom under the supervision of their class teachers. The scoring sheets of all tests were collected after the testing session, data were centrally processed, and no feedback was provided to the children.

The online data collection was carried out via the *eDia* (Electronic Diagnostic Assessment) platform through the Internet. Testing took place in the computer labs at the participating schools, using the available computers and browsers installed. A session lasted approximately 20–45 min, depending on the test. The items were automatically scored, and children received immediate feedback (percent of correct answers) at the end of the testing session. The *eDia* platform allows the use of proxy servers.

Ezen a képen egy ház és négy madár látható. Mutasd meg, melyik madár van fenn!



Figure 1. Sample item from the computer-based version of the relational reasoning test. [In this picture, you can see a house and four birds. Click on the bird that is higher up than the others.]



Kattints rá arra, amelyikben 1 rajz van!

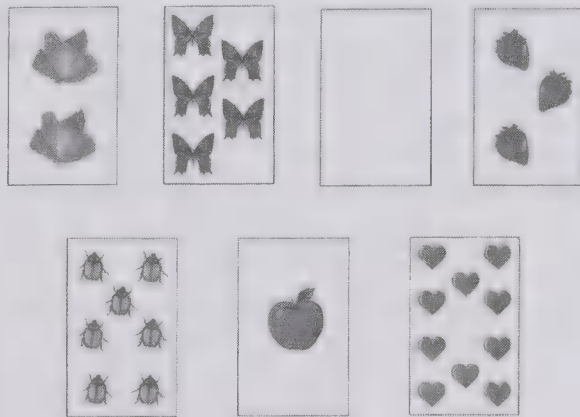


Figure 2. Sample item from the computer-based version of the counting and basic numeracy skills test. [Click on the card with a drawing of only one thing on it.]

### Statistical Analyses

In this article, we analyze the differences between paper-based, FF, and online tests. We not only examine whether tests presented in different modes are equivalent, but we also show where they are different, as one of the aims of this study was to support the design of better online instruments. To reach this goal, we applied several analyses, including computations based on classical test theory, confirmatory factor analyses within structural equation modeling (SEM; Bollen, 1989), to test the underlying measurement model and to test measurement invariance, and IRT. In this section, we only discuss the theoretical background of how SEM analyses were applied in the present study.

Providing a meaningful interpretation of test scores and ensuring the comparability and validity of FF and CB test results is only possible if the structure of the construct does not change across delivery modes (Byrne & Stewart, 2006). That is, measurement invariance must be analyzed to examine whether test results are affected by the test medium and to ensure that the same constructs are being assessed in each group. If measurement invariance is sufficiently met, and, thus, structural stability exists, between-group differences can be interpreted as true and not as psychometric differences in latent ability (Greiff et al., 2013).

A number of approaches, statistical methods, and concepts are available to test measurement equivalence (Schroeders & Wilhelm, 2011). State-of-the-art methods share a common feature: The definition of the measurement model is provided through a comparison of the latent structure for several groups in a single model. The most prominent methods are those used to detect differential item functioning within the IRT approach (Raju, Laffitte, & Byrne, 2002) and multigroup confirmatory factor analysis (MGCFA) (Bollen & Curran 2006; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000) within the SEM framework.

In the present study, a between-subject design was used to test invariance by means of MGCFA. Weighted least squares, mean- and variance-adjusted (WLSMV) estimation was applied, and THETA parameterization was used because all items were scored dichotomously (Muthén & Muthén, 2010). All measurement mod-

els were computed with Mplus. Goodness of fit to the sample data was evaluated on the basis of multiple criteria. Different fit indices have been developed (Wu, Li, & Zumbo, 2007), and numerous cutoff criteria, such as the Tucker–Lewis Index (TLI), comparative fit index (CFI)  $\geq 0.90$  or  $0.95$ , and root-mean-square error of approximation (RMSEA)  $\leq 0.06$  or  $0.08$ , have been proposed to assist in determining model fit (see Byrne & Stewart, 2006; Fan & Sivo, 2005; Vandenberg & Lance, 2000). In this study, an absolute fit index (the RMSEA), a relative fit index (the TLI), and an incremental, normed fit index (the CFI) were used to evaluate model fit. Nested model comparisons were conducted using a special chi-square difference test for the WLSMV estimator (Muthén & Muthén, 2010).

Testing for measurement invariance (Muthén & Muthén, 2010; Vandenberg & Lance, 2000) with categorical data involves a fixed sequence of model comparisons, testing different levels of invariance by comparing measurement models from the least to the most restrictive model by using MGCFA. Measurement invariance exists if restrictions of model parameters in one model do not generate a substantially worse model fit in comparison to an unrestricted model. The procedure for testing measurement invariance is explained thoroughly by Byrne and Stewart (2006).

Configural invariance investigates whether the basic model structure is invariant across groups (Byrne, 2008), that is, whether children in the CB and FF environments conceptualize the construct in the same way (Milfont & Fischer, 2010) and thus use the same conceptual framework to answer the test items (Vandenberg & Lance, 2000; Wu, Li, & Zumbo, 2007). Configural invariance indicates that the same item is an indicator of the same latent factor in each group, but factor loadings may differ across groups. When we tested configural invariance with categorical outcomes, thresholds and factor loadings were not constrained across groups, factor means were fixed at 0 in all groups, and residual variances were fixed at 1 in all groups. The next step is to test weak factorial invariance, that is, to test cross-group equality in the loadings. However, testing it for categorical data is not recommended (Muthén & Muthén, 2010), and thus weak factorial invariance was not tested (see, e.g., Greiff et al., 2013; Schroeders & Wilhelm, 2011).

Strong invariance, as the subsequent step in testing measurement invariance, indicates that the variances for latent variables and the covariances between the latent variables are equal between

Kattints rá arra a három alakzatra, amelyekben van valami közös és különböznek ■ többletöl



Figure 3. Sample item from the computer-based version of the inductive reasoning test. [Click on the three shapes that have one thing in common that the other two do not.]



CB and FF modes; that is, cross-group equality exists in the loadings and intercepts. When this was tested, thresholds and factor loadings were constrained so that they would be equal across groups, and residual variances were fixed at 1 and factor means at 0 in the FF group, whereas there were no constraints specified in the CB group (Muthén & Muthén, 2010). If strong factorial invariance did not hold according to the modification indices, partial strong invariance was tested. Strong factorial invariance is the level at which latent mean comparisons can be conducted (Byrne & Stewart, 2006).

Finally, strict factorial invariance indicates whether the CB and FF groups have the same item residual variances (Byrne, 2008). It requires cross-group equality in the loadings, intercepts, and residual variances. Therefore, in addition to the restrictions applied in strong factorial invariance, all residual variances were fixed at one in all groups, even though strict factorial invariance is not a prerequisite for media comparisons of latent factor means and variances.

## Results

The presentation of the results is organized according to the research questions. First, we examine measurement invariance for each of the five tests (Research Question 1). Second, as we see that in some cases the scales did not remain identical, we study the direction of the changes (Research Question 2). Finally, we examine whether changing the testing media has the same impact on boys and girls (Research Question 3).

### Research Question 1: Examining the Media Effect Through Analyses of Measurement Invariance

The measurement invariance analyses were performed as described in the Method section. The results are summarized in Table 2.

For the speech sound discrimination test, both the FF and the CB versions were administered to the same sample, and a multivariate, single-level approach made it possible to test measurement invariance in a single-group analysis. First, we tested the confirmatory factor analysis at each of the two points in time to be sure the model fit well in both modes. Examination of the modification indices suggested that model fit would be significantly improved by changing the original model. According to the results of the LaGrange multiplier test, we needed to delete some 16 items from the analyses because of ceiling effects. The remaining 44 items fit the data in both modalities well (FF: RMSEA = .019, CFI = .943, TLI = .940; CB: RMSEA = .035, CFI = .915, TLI = .910). The strong factorial invariance model did not fit well and resulted in a significant decrease in fit relative to the configural invariance model. The examination of the modification indices suggested that model fit would be significantly improved by allowing the intercept for one item to differ between data collections and adding residual covariances between two items to the CB model. Partial strong invariance did hold. The observed differences in item means between PP and CB testing was due to factor mean differences (for one item, children in FF mode were expected to have higher item response) and a residual covariance in CB mode (two items proved to be correlated). Finally, we tested partial strict invariance, resulting in a significant decrease in fit relative to the configural model; that is, the relations of the items to the latent factor of speech sound discrimination were not equivalent in PP and CB modes in a strict sense. However, strict factorial invariance is not a prerequisite for latent mean comparisons; in this case, partial strong factorial invariance is sufficient to compare latent means.

The results regarding the strong factorial invariance model for relational reasoning indicated a significant decrease in fit relative to the configural invariance model. The modification indices suggested a freeing of the intercept for two items between PP and CB

Table 2  
*Goodness-of-Fit Indices for Measurement Invariance of the Tests*

Test	Model	$\chi^2$	df	CFI	TLI	RMSEA	$\Delta\chi^2$ <sup>a</sup>	$\Delta df$ <sup>a</sup>	p
Speech sound discrimination	(1)	2905.7	2628	.908	.906	.017			
	(2)	2950.2	2663	.905	.904	.017	60.0	35	<.05
	(2.1)	2935.9	2661	.909	.908	.017	44.8	33	>.05
	(3)	3231.4	2700	.824	.824	.024	287.0	72	<.05
Relational reasoning	(1)	389.9	97	.971	.961	.051			
	(2)	490.5	108	.962	.954	.055	89.2	11	<.01
	(2.1)	396.0	105	.971	.964	.049	27.4	8	<.01
	(3)	1027.1	119	.910	.900	.081	463.9	22	<.01
Counting and numeracy	(1)	46.2	17	.996	.997	.038			
	(2)	243.4	22	.984	.978	.093	128.2	5	<.01
	(2.1)	157.0	20	.990	.985	.076	70.8	3	<.01
	(3)	268.9	27	.983	.981	.087	169.6	10	<.01
Deductive reasoning	(1)	185.8	136	.980	.973	.030			
	(2)	293.1	146	.942	.927	.049	75.9	10	<.01
	(2.1)	264.4	142	.951	.938	.046	56.4	6	<.01
	(3)	348.1	166	.928	.921	.051	34.0	14	<.01
Inductive reasoning	(1)	1791.2	908	.929	.923	.037			
	(2)	1828.4	930	.924	.919	.038	42.0	22	>.01
	(2.1)	1806.4	931	.926	.921	.038	15	23	>.05
	(3)	1868.1	962	.921	.916	.039	55.0	32	>.05

Note. Model: (1) = configural invariance; (2) = strong factorial invariance; (2.1) = partial strong factorial invariance; (3) = strict factorial invariance. CFI = comparative fit index; TLI = Tucker–Lewis Index; RMSEA = root-mean-square error of approximation.

<sup>a</sup>  $\Delta\chi^2$  and  $\Delta df$  were estimated with the Difference Test procedure (DIFFTEST) in Mplus. When using weighted least squares, mean- and variance-adjusted estimation,  $\chi^2$  differences between models cannot be compared by subtracting  $\chi^2$  and  $df$  (Muthén & Muthén, 2010).



testing. The partial strong invariance model in which the intercept for these two items was allowed to differ between PP and CB groups fit better than the strong factorial invariance model, but still significantly worse than the configural invariance model. This means no (partial) measurement invariance could be established for the relational reasoning test.

For counting and basic numeracy, strong factorial invariance model resulted in a significant decrease in fit relative to the configural invariance model as well. According to the LaGrange multiplier test, the intercept for Items 6 and 7 between PP group and CB group were freed. The partial strong invariance model fit better than strong factorial invariance model, but still significantly worse than configural invariance model. Thus, no measurement invariance could be established in this case either.

The result for the invariance testing of deductive reasoning indicated a decrease in model fit for all levels of invariance; thus, there was no measurement or partial measurement invariance between the FF and CB testing modes for deductive reasoning. These data suggest that CB (online administration) does not measure exactly the same construct as FF delivery does. To this end, mean differences between FF and CB groups could not be interpreted as true differences in the underlying deductive reasoning construct; this could also be due to psychometric issues. One of the possible reasons for this is that information was more standardized in the CB environment; achievements in the CB environment were independent of the teacher's attitude and judgment during data collection. This was not the case in the FF mode.

The result for the invariance testing of inductive reasoning represented no loss in model fit; even imposing the most restrictive constraints did not lead to deterioration in model fit. The model of strong factorial invariance did not show a decrease in model fit compared with the model of configural invariance. Strict factorial invariance could also be established; that is, even residual variances proved to be equal across delivery media in a strict sense.

## Research Question 2: Differences Between the Tests Delivered in Different Media

As shown in the previous section, depending on the content of the assessment and the item types, there are differences between the tests in respect of how the measurement scales changed when the items were transferred to the online platform. In this section, we compare the reliability of the tests delivered by the two media, examine the impact of the media on performance, and have a closer look at the item level

differences by comparing the difficulties of the items in the two media.

**Reliability of the tests.** The internal consistencies of the tests were examined by computing Cronbach's alpha for each test. The DIFER tests were previously also administered to participants in the Hungarian Educational Longitudinal Program ( $5,000 > n > 6,000$ ), and the reliability indices of those two digitized in this study were high (relational reasoning: .726; counting and basic numeracy: .915), the relational reasoning test showing the lowest value (Csapó, 2007; Józsa, 2004). In the present study, the reliability indices were slightly different but in general also good both in FF/PP and CB modes. They ranged from .743 to .887 in the FF/PP mode and from .770 to .938 in the CB mode. Generally, the reliability indices of the CB tests proved to be somewhat higher than those of the FF/PP test versions (see Table 1).

The reliability value was already high for the FF administration of speech sound discrimination (.887), and it improved further (to .938), being the highest within this set of tests. This improvement may be attributed to the standardized voice stimuli. There were slight improvements in Cronbach's alpha for relational reasoning and deductive reasoning. A major drop of reliability was observed for the counting and basic numeracy test. Although several items were dropped from the FF version because they required oral responses and it was not possible to implement this in the CB version, the reduced FF test consisting of 13 items still had a relatively high reliability (.813). The PP version of the inductive reasoning test was digitized without major changes. This was reflected in the reliabilities as Cronbach's alphas of the two versions did not differ.

**The impact of the assessment media on performance.** A meaningful interpretation of differences in test scores is only possible if the structure of the construct measured does not change across test media (Byrne & Stewart, 2006). That is, latent (and manifest) mean comparison can only be interpreted meaningfully if at least strong factorial invariance is established (Brown, 2006). According to the measurement invariance analyses, testing for latent mean differences was only meaningful in the case of speech sound discrimination and inductive reasoning tests, where partial strong and strict measurement invariance held, respectively.

Latent mean comparisons were conducted by constraining the item intercepts of the observed variables equal and setting the latent factor means for the FF (or PP) group as reference group to zero (Byrne & Stewart, 2006). We also calculated performance on tests in percentages, summarizing the results in Table 3. We report

Table 3  
Test-Level Achievement Differences Between Traditional (FF or PP) and CB Modes

Test	FF or PP (%)		CB (%)		<i>d</i>	Latent		<i>p</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		<i>M</i>	<i>SE</i>	
Speech sound discrimination	91.35	9.88	82.61	16.80	.59	-.77	.13	<.01
Relational reasoning <sup>a</sup>	80.86	15.44	78.03	17.86	.17	-.67	.11	<.01
Counting and basic numeracy <sup>a</sup>	87.35	18.46	88.90	14.84	-.09	.19	.07	<i>ns</i>
Deductive reasoning <sup>a</sup>	70.69	14.34	63.59	20.87	.39	-.58	.06	<.01
Inductive reasoning	47.52	19.04	45.99	18.11	.08	-.10	.06	<i>ns</i>

Note. *d* is the difference between the means in standard deviation units (Cohen's *d*). Latent mean for the FF group was set to zero. FF = face-to-face; PP = paper and pencil; CB = computer based.

<sup>a</sup> Measurement invariance does not hold.

the means also for those tests for which measurement invariance could not be established.

The data indicate that performance on the speech sound discrimination test was significantly lower in the CB tests than in the FF test: it fell from a high (91.4%) to a still high but significantly lower level (82.6%). A similar drop was found in the case of deductive reasoning and a modest drop for relational reasoning. No significant decreases in mean differences were observed for the counting and basic numeracy test or the inductive reasoning test. The latter represents a PP-CB transition, and no significant difference was found between the two versions.

**Item-level differences.** A further way of analyzing the characteristics of CB testing is to have a look at the results at the item level. To do this, we computed the item difficulties for each item in both media. To illustrate the possibilities for this type of analysis, we present the results of the speech sound discrimination, deductive reasoning, and inductive reasoning tests.

For the speech sound discrimination test, we once again took advantage of the two test versions being administered to the same sample and computed the item parameters on the basis of IRT scaling. We considered the entire item pool of the two versions of the test as items of a single test and calculated the item parameters. According to the analysis of the content, items that contained nonsense words proved to be most affected by FF administration. A possible reason for this difference may be that in the case of meaningless words, teachers helped children to find the correct answer. The same effect was observed for words where the lack of context could in part have been replaced by teachers' helpful behavior, whereas the impact of FF administration was less apparent for words in complete sentences (only in the case of this latter test was a significant correlation found between the item difficulty parameters:  $r = .550, p < .01$ ). These results also suggest that items presented by teachers lose their objectivity in some cases, especially if stimuli are taken out of their usual context.

For the deductive reasoning tests, a much higher correlation was found between the item difficulties in the two media ( $r = .750, p < .01$ ). The inductive reasoning test showed the most "regular" picture, in agreement with the previous observations. The items correlated to a very high degree ( $r = .948, p < .01$ ), indicating that the PP and CB tests measure inductive reasoning skills very similarly, not only at the overall test level but also at the level of items as well. The inductive reasoning test is the only one where one of the versions was taken on paper. The same can be observed here as what we have already shown concerning the reliability and the difference between the means: The two versions of the test behave very much alike, so they can essentially be considered identical.

### Research Question 3: Gender Differences

As previous studies have indicated, gender differences are influenced by the medium of testing. In general, boys perform somewhat better if they are assessed via technology-based instruments. In the previous FF versions of the DIFER tests, girls performed somewhat better on the speech sound discrimination test, whereas boys performed better on the counting and numeracy test (Józsa, 2004).

To examine how transferring the DIFER tests and the inductive reasoning test to the online platform affected gender differences at

this very young age, we carried out several analyses. First, we performed invariance analyses with regard to gender.

The model used to test configural invariance of speech sound discrimination for boys and girls fit well. The model of strong factorial invariance did not show a decrease in model fit based on the stricter perspective (nonsignificant chi-square difference test; cf. Table 4) compared with the model of configural invariance. Finally, strict factorial invariance could not be established. As strong factorial invariance held (see Meredith, 1993), mean differences could be interpreted as true differences in the construct being measured between girls and boys (Byrne & Stewart, 2006).

In the computerized versions of the tests examined here, a significant gender difference was only found for the speech sound discrimination test: Girls performed somewhat better than boys (in %, girls:  $M = 86.37, SD = 12.76$ ; boys:  $M = 80.40, SD = 18.03$ ,  $t = -3.94, p < .001$ ). No significant gender differences were found for the other tests.

## Discussion

### Measurement Invariance

We found that measurement invariance held in two out of the five cases (if we consider the practical perspective and the strong invariance model sufficient) and, in a strict sense, only in one case, for the inductive reasoning test. This test was originally a PP test, and the migration of the items took place so that neither the general look of the test nor the item types were changed.

Measurement invariance held only partially for the speech sound discrimination test. This was originally an individually administered FF test, and the same pictures were presented to the children in both modes. The item type and the scoring principle did not change either. This finding indicates that under certain conditions, even an FF test can be transferred to the online platform with acceptable results in terms of measurement invariance. This may also hold in part for the relational reasoning test.

Measurement invariance did not hold for the three remaining tests. These results indicate that equivalent scales may only be constructed if the migration of the items does not change the item types and only changes the testing context moderately. If the migration influences the objectivity of scoring, the tests administered in the two media are not exactly identical. This happened on

Table 4  
*Goodness-of-Fit Indices for Measurement Invariance of the Speech Sound Discrimination Test for Boys and Girls*

Model	$\chi^2$	df	CFI	TLI	RMSEA	$\Delta\chi^2$ <sup>a</sup>	$\Delta df$ <sup>a</sup>	p
(1)	2858.7	2536	.932	.929	.024			
(2)	2887.0	2580	.935	.934	.024	44.8	44	>.05
(3)	2920.2	2588	.930	.928	.024	80.2	52	<.05

*Note.* Model: (1) = configural invariance; (2) = strong factorial invariance; (3) = strict factorial invariance. CFI = comparative fit index; TLI = Tucker-Lewis Index; RMSEA = root-mean-square error of approximation.

<sup>a</sup>  $\Delta\chi^2$  and  $\Delta df$  were estimated with the Difference Test procedure (DIFFTEST) in Mplus. When using weighted least squares, mean- and variance-adjusted estimation,  $\chi^2$  differences between models cannot be compared by subtracting  $\chi^2$  and  $df$  (Muthén & Muthén, 2010).



the deductive reasoning test as the open-ended items were converted to multiple-choice items.

From an applied perspective, the scale differences may only cause problems if there is a need to compare performance within the two media. If the purpose of transferring the test is to construct a new, more applicable instrument, the lack of measurement invariance does not cause a problem. As the aim of the present study was to establish the development of a new instrument, analyzing further details of the differences where the measurement invariance does not hold may help to construct better tests. If the changes may be attributed to the improvement of the instrument, the lack of invariance may even be favorable, but in this case, the new instruments can only be used in practice after careful piloting and validation processes.

### **The Impact of Media on Reliability and Achievement Scores**

The study has shown that the reliability of assessments may be improved by transferring individually administered FF instruments to an online platform. Having a look at the reliability coefficients of the five instruments in two modes, we may observe that the reliability increased in those cases when CB assessment provided more standardized conditions compared with FF administration.

The results indicated that performance was lower on the CB tests than on the FF or PP tests in most cases. This suggests that teachers tend to give higher scores to children than the scores they receive when their responses are automatically (and objectively) scored. Another explanation for the differences might be that children had difficulty handling the computerized tests, and this lowered their performance. However, as no significant mean differences were found for the counting and basic numeracy test or the inductive reasoning test, low computer familiarity is not a sufficient explanation for achievement differences. In fact, a more realistic explanation may be that teachers were more tolerant in accepting children's responses.

### **The Impact of Media on Gender Differences**

Due to the limitation of the available data, gender differences were explored by latent analyses only for the speech sound discrimination test. The results indicated that measurement invariance did hold; thus, FF and CB testing was the same for boys and girls, and latent means could be compared. Transferring the test to the new medium affected their performance similarly. A comparison of the raw scores indicated that girls' performance was somewhat better, similarly to former FF assessments, where girls usually performed slightly better than boys. Together, these data may indicate that using the new medium will not cause a major bias in future applications.

### **Limitations of the Present Study**

Due to the context of the study (using a system that is still under development), there were smaller sample sizes available compared with previous large-scale assessments. The availability of computers at school and the time first-graders were available to work on the computers limited the possible complexity of the study. The analyses are also constrained by the unavailability of additional background variables. In this phase of the research, we have no data concerning the possibilities of using the same instrument for

repeated testing to monitor development. (The previous FF version of the DIFER is routinely used for this purpose without problems.) There are no data on the predictive validity of the CB instruments either. However, as indicated earlier, the FF version was administered in a longitudinal study and proved to be a good predictor of later school performance.

A further limitation of the present study is that the original DIFER tests were designed to assess children in the kindergarten-to-school transition period. Thus, students who have already started school tend to perform close to ceiling on these tests, and their data are not ideal for analyzing the characteristics of the tests. This deficiency may be rectified by extending future investigations to kindergarten populations, although the use of computers with that age group calls for further feasibility studies.

### **Conclusions and Further Prospects for Online Assessment of School Readiness**

This study shows the potential and limitations of transferring school readiness tests to the new assessment medium of computers. In the digitization process, we have lost a strong and important test with high reliability and predictive validity as the social skills test cannot be replicated in the new medium while remaining close to its original form. We have also lost the writing test, but a closely related construct (fine hand movement) can easily be measured using emerging technologies. In fact, an alternative construct (handling keyboard and mouse) can also be easily measured by computer. The relevant technology is at hand for research purposes, and it will probably also be widely available in schools. We have also lost a large number of relevant items on the counting and basic numeracy skills test, as it was not possible to capture children's oral responses, and we have paid for this loss with a drop in reliability. Further efforts are therefore needed to develop a suitable CB counting skills test.

The study has demonstrated the applicability of technology-based assessment in regular school practice at the earliest possible point of schooling in a number of highly relevant competency domains. These assessments can be carried out practically anytime, at very low cost, and with almost no extra teacher time. However, devising and using such instruments require further research in at least three dimensions: (a) making constructs currently assessed with traditional instruments measurable using computer technology and assessing new constructs that are especially well suited to CB assessment; (b) enhancing the online assessment technology with functionalities that are already in use in other areas of information technology (e.g., speech recognition and detection of emotions); and (c) exploring ways of integrating frequent early assessment into educational processes.

A number of technological solutions that can be built into the online assessment system to enhance its capabilities already exist elsewhere. Interaction, simulation, and manipulation of objects on screen, and new types of stimuli, such as video and animation, are currently in use in some assessments. It is also possible to time the stimuli and control the presentation of information in other ways. Measuring response time, logging keystrokes, and mouse movement can also routinely be used, although further studies are required to explore how these methods may contribute to solving the real problems of early assessment. One example of a real problem, where an existing technological solution may be essen-



tial, emerges from the present study: Voice recognition technology is needed to make the counting test deliverable online.

Further research is needed to explore the educational applicability of online assessment. Ecological validity is an issue that requires careful consideration. Examining predictive validity is crucial for tests that assess the preconditions of further learning and are used to identify early indicators of later problems. An examination of the online tests discussed here has already started as an extension of the present study. An exploration of the effects of repeated testing has also begun, but the accumulation of a sufficient quantity of data for analyses will take time in both cases. As learners' assessment results can easily be stored in the online assessment system, it is possible to gather not only overall performance data but also information collected on behavioral processes. In addition, the growing information base facilitates adequate monitoring of learners' development.

In the past decade, the issue of early development has been approached not only by researchers in numerous fields of study in education and psychology but also by those in other social sciences, such as sociology and economics. Results from these comprehensive studies have indicated that numerous problems that arise later are rooted in difficulties in the first school years. Research has also shown that these difficulties may be overcome with adequate intervention and that investing in such programs produces high returns. An important component of the well-prepared and well-timed intervention is proper diagnosis. To this end, CB assessment of the basic skills may be one of the best means to diagnose problems and monitor development.

## References

- Blair, C. (2002). School readiness: Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry. *American Psychologist*, 57, 111–127. doi:10.1037/0003-066X.57.2.111
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.
- Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley.
- Bossaert, G., Doumen, S., Buyse, E., & Verschueren, K. (2011). Predicting children's academic achievement after the transition to first grade: A two-year longitudinal study. *Journal of Applied Developmental Psychology*, 32, 47–57. doi:10.1016/j.appdev.2010.12.002
- Brown, T. (2006). CFA with equality constraints, multiple groups, and mean structures. In T. Brown (Ed.), *Confirmatory factor analysis for applied research* (pp. 236–319). New York, NY: Guilford Press.
- Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema*, 20, 872–882.
- Byrne, B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling*, 13, 287–321. doi:10.1207/s15328007sem1302\_7
- Carlton, M. P., & Winsler, A. (1999). School readiness: The need for a paradigm shift. *School Psychology Review*, 28, 338–352.
- Carson, K., Gillon, G., & Boustead, T. (2011). Computer-administrated versus paper-based assessment of school-entry phonological awareness ability. *Asia Pacific Journal of Speech, Language and Hearing*, 14, 85–101.
- Choi, S. W., & Tinkler, T. (2002). *Evaluating comparability of paper and computer based assessment in a K-12 setting*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology*, 33, 593–602. doi:10.1111/1467-8535.00294
- Csapó, B. (1997). The development of inductive reasoning: Cross-sectional assessments in educational context. *International Journal of Behavioral Development*, 20, 609–626. doi:10.1080/016502597385081
- Csapó, B. (2007). Hosszmetszeti felmérések iskolai kontextusban - az első átfogó magyar iskolai longitudinális kutatási program elméleti és módszertani keretei [Longitudinal assessments in school context – theoretical and methodological frameworks of the first large-scale school-related longitudinal program in Hungary]. *Magyar Pedagógia*, 107, 321–355.
- Csapó, B. (2013, May). *The predictive validity of school readiness assessment: Results from an eight-year longitudinal study*. Poster presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Csapó, B., Ainley, J., Bennett, R., Latour, T., & Law, N. (2012). Technological issues of computer-based assessment of 21st century skills. In B. McGaw, P. Griffin, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 143–230). New York, NY: Springer. doi:10.1007/978-94-007-2324-5\_4
- Csapó, B., Molnár, G., & Tóth, K. R. (2009). Comparing paper-and-pencil and online assessment of reasoning skills: A pilot study for introducing TAO in large-scale assessment in Hungary. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 113–118). Luxemburg, Belgium: Office for Official Publications of the European Communities.
- Csapó, B., & Nikolov, M. (2009). The cognitive contribution to the development of proficiency in a foreign language. *Learning and Individual Differences*, 19, 209–218. doi:10.1016/j.lindif.2009.01.002
- Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, 43, 1428–1446. doi:10.1037/0012-1649.43.6.1428
- Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling*, 12, 343–367. doi:10.1207/s15328007sem1203\_1
- Gallagher, A., Bridgeman, B., & Cahalan, C. (2000). *The effect of computer-based tests on racial/ethnic, gender, and language groups* (GRE Board Professional Report No. 96-21P). Princeton, NJ: Education Testing Service.
- Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement*, 36, 189–213. doi:10.1177/0146621612439620
- Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F., & Funke, J. (2013). Computer-based assessment of complex problem solving: Concept, implementation, and application. *Educational Technology Research and Development*, 61, 407–421. doi:10.1007/s11423-013-9301-x
- Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts—Something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology*, 105, 364–379. doi:10.1037/a0031856
- Guhn, M., Janus, M., & Hertzman, C. (2007). The Early Development Instrument: Translating school readiness assessment into community actions and policy planning. *Early Education & Development*, 18, 369–374. doi:10.1080/10409280701610622
- Hair, E., Halle, T., Terry-Humen, E., Lavelle, B., & Calkins, J. (2006). Children's school readiness in the ECLS-K: Predictions to academic, health, and social outcomes in first grade. *Early Childhood Research Quarterly*, 21, 431–454. doi:10.1016/j.ecresq.2006.09.005
- Horne, J. (2007). Gender differences in computerised and conventional educational tests. *Journal of Computer Assisted Learning*, 23, 47–55. doi:10.1111/j.1365-2729.2007.00198.x



- Józsa, K. (2004). Az első osztályos tanulók elemi alapképességeinek fejlettsége Egy longitudinális kutatás első mérési pontja [Developmental level of first-grade students' basic skills. The first measurement point of a longitudinal research program]. *Iskolakultúra*, 14, 3–16.
- Kingston, N. M. (2008). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education*, 22, 22–37. doi:10.1080/08957340802558326
- Klauer, K. J. (1989). *Denktraining für Kinder I* [Training of thinking for children]. Göttingen, Germany: Hogrefe.
- Konold, T. R., & Pianta, R. C. (2005). Empirically-derived, person-oriented patterns of school readiness in typically-developing children: Description and prediction to first-grade achievement. *Applied Developmental Science*, 9, 174–187. doi:10.1207/s1532480xads0904\_1
- Magnuson, K. A., Ruhm, C., & Waldfogel, J. (2007). The persistence of preschool effects: Do subsequent classroom experiences matter? *Early Childhood Research Quarterly*, 22, 18–38. doi:10.1016/j.ecresq.2006.10.002
- Mashburn, A. J., & Henry, G. T. (2004). Assessing school readiness: Validity and bias in preschool and kindergarten teachers' ratings. *Educational Measurement: Issues and Practice*, 23, 16–30. doi:10.1111/j.1745-3992.2004.tb00165.x
- McWayne, C. M., Cheung, K., Wright, L. E. G., & Hahs-Vaughn, D. L. (2012). Patterns of school readiness among head start children: Meaningful within-group variability during the transition to kindergarten. *Journal of Educational Psychology*, 104, 862–878. doi:10.1037/a0028884
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525–543. doi:10.1007/BF02294825
- Merrell, C., & Bailey, K. (2012). Predicting achievement in the early years: How influential is personal, social and emotional development? *Online Educational Research Journal*. Retrieved from <http://www.oerj.org/View?action=viewPaper&paper=55>
- Merrell, C., & Tymms, P. (2011). Changes in children's cognitive development at the start of school in England 2001–2008. *Oxford Review of Education*, 37, 333–345. doi:10.1080/03054985.2010.527731
- Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, 3, 111–121.
- Molnár, G. (2011). Playful fostering of 6- to 8-year-old students' inductive reasoning. *Thinking Skills and Creativity*, 6, 91–99. doi:10.1016/j.tsc.2011.05.002
- Molnár, G., & Csapó, B. (2011). Az 1–11 évfolyamot átfogó induktív gondolkodás kompetenciaskála készítése a valószínűségi tesztelmélet alkalmazásával [Constructing inductive reasoning competency scales for years 1–11 using IRT models]. *Magyar Pedagógia*, 111, 127–140.
- Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.
- Nagy, J. (1980). *5–6 éves gyermekeink iskolakészültsége* [School readiness among 5- to 6-year-old children]. Budapest, Hungary: Akadémiai Kiadó.
- Nagy, J. (1987). *Prefer: Preventív fejlettségvizsgáló rendszer 4–7 éves gyermekek számára* [A test battery for assessment of 4- to 7-year-old children's school entry competencies]. Budapest, Hungary: Akadémiai Kiadó.
- Nagy, J., Józsa, K., Vidákovich, T., & Fazekasé Fenyvesi, M., (2004a). Az elemi alapképességek fejlődése 4–8 éves életkorban. Az eredményes iskolakezdés hét kritikus alapképességének országos helyzetképe és a pedagógiai tanulságok [The development of elementary skills between the ages of 4 and 8. A national overview of the seven basic skills needed for academic success and their pedagogic consequences]. Szeged, Hungary: Mozaik Kiadó.
- Nagy, J., Józsa, K., Vidákovich, T., & Fazekasé Fenyvesi, M. (2004b). *Diagnosztikus fejlődésvizsgáló és kritériumorientált fejlesztő rendszer 4–8 évesek számára: DIFER programcsomag* [Diagnostic assessment and criterion-oriented development system for 4- to 8-year-olds: The DIFER package]. Szeged, Hungary: Mozaik Kiadó.
- Organisation for Economic Co-Operation and Development. (2010). *PISA computer-based assessment of student skills in science*. Paris, France: Author.
- Organisation for Economic Co-Operation and Development. (2011). *PISA 2009 results: Students on line: Digital technologies and performance (Volume VI)*. Paris, France: Author.
- Organisation for Economic Co-Operation and Development. (2014). *Skills for life: Student performance in problem solving*. Paris, France: Author.
- Price, P., Tepperman, J., Iseli, M., Duong, T., Black, M., Wang, S., . . . Alwan, A. (2009). Assessment of emerging reading skills in young native speakers and language learners. *Speech Communication*, 51, 968–984. doi:10.1016/j.specom.2009.05.001
- Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87, 517–529. doi:10.1037/0021-9010.87.3.517
- Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement*, 71, 849–869. doi:10.1177/0013164410391468
- Shepard, L. A. (1997). Children not ready to learn? The invalidity of school readiness testing. *Psychology in the Schools*, 34, 85–97. doi:10.1002/(SICI)1520-6807(199704)34:2<85::AID-PITS2>3.0.CO;2-R
- Snow, C. E., & Van Hemel, S. B. (Eds.). (2008). *Early childhood assessment: Why, what, and how*. Washington, DC: National Academies Press.
- Snow, K. L. (2006). Measuring school readiness: Conceptual and practical considerations. *Early Education and Development*, 17, 7–41. doi:10.1207/s15566935eed1701\_2
- Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25, 78–107. doi:10.1086/209528
- Tymms, P., Jones, P., Albane, S., & Henderson, B. (2009). The first seven years at school. *Educational Assessment, Evaluation and Accountability*, 21, 67–80. doi:10.1007/s11092-008-9066-7
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the MI literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4–70. doi:10.1177/109442810031002
- Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 assessment: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, 68, 5–24.
- Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). *Score comparability of online and paper administrations of Texas assessment of knowledge and skills*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation*, 12(3). Retrieved from <http://pareonline.net/pdf/v12n3.pdf>

Received April 18, 2013

Revision received December 9, 2013

Accepted December 27, 2013 ■

# Toward Automated Computer-Based Visualization and Assessment of Team-Based Performance

Dirk Ifenthaler  
Deakin University

A considerable amount of research has been undertaken to provide insights into the valid assessment of team performance. However, in many settings, manual and therefore labor-intensive assessment instruments for team performance have limitations. Therefore, automated assessment instruments enable more flexible and detailed insights into the complex processes influencing team performance. The central objective of this study was to advance knowledge in automated assessment of team-based performance using a language-oriented approach. Fifty-six teams of learners ( $N = 224$ ) in 3 experimental conditions solved 2 tasks in an online learning environment. They were analyzed with the Automated Knowledge Visualization and Assessment (AKOVIA) methodology. AKOVIA integrates a natural language-oriented algorithm and enables a structural and semantic compression of individual- and team-based knowledge representations. Findings indicate initial evidence of the feasibility and validity of the fully automated methodology. A framework for integrating research and methodology development is suggested for improving educational technology innovations such as computer-based assessment environments in international large-scale assessments.

**Keywords:** team, shared mental model, automated assessment, natural language processing

Teams are a critical and essential part of most working environments because they combine different views, multiple skills, diverse experiences, analytical judgments, and rich knowledge. Consequently, research in teams and their assessment has been a continuous endeavor in various scientific areas for more than 30 years. Yet, there exist various definitions of *team* using different perspectives. For example, Kanaga and Kossler (2011) defined a team as “a specific kind of group whose members are collectively accountable for achieving the team’s goals” (p. 4). A more detailed definition is given by Katzenbach and Smith (2003), who described a team as “a small number of people with complementary skills who are committed to a common purpose, performance goals, and approach for which they are mutually accountable” (p. 45). From an operational point of view, Cohen, Levesque, and Smith (1997) defined a team as “a set of agents having a shared objective and a shared mental state” (p. 95). Salas, Dickinson, Converse, and Tannenbaum (1992) described a team as

a distinguishable set of two or more people who interact dynamically, interdependently, and adaptively toward a common and valued goal, who have each been assigned specific roles or functions to perform and who have a limited life span of membership. (p. 4)

To sum up, common characteristics of definitions of a team include at least two individuals, common objectives, shared responsibility and interdependence, as well as optimal performance.

Instruments for measuring team performance have been developed over the past decades; however, adequate computer-based assessments of team-based performance are scarce (Fischer & Mandl, 2005). The recent advancement of web-based technology allowed widening the scope of computer-based assessments (Csapó, Ainley, Bennett, Latour, & Law, 2012; Frey & Hartig, 2013). For example, international large-scale assessments such as the Programme for International Student Assessment (PISA) or the Programme for the International Assessment of Adult Competencies (PIAAC) currently implement advanced computer-based assessment environments (Organisation for Economic Co-operation and Development [OECD], 2010, 2013).

Previously, most of the team-based assessment instruments required a great deal of time and effort using highly trained researchers (e.g., think-aloud protocol analysis) and were mainly limited to subjective self-reports (Wildman et al., 2012), and they also required labor-intensive manual analysis of performance indicators (Almond, Steinberg, & Mislevy, 2002). As a result, such assessments have been limited to the scientific community and have had only a minor impact on practical issues such as the design of effective learning, teaching, and working environments. Motivated by a desire to have practical assessment instruments that are useful and valid has led researchers to uncover significant developments in the last several years (Chung, O’Neil, & Herl, 1999; Mandl & Fischer, 2000). Especially instruments using graphical representations for computer-based assessment have been successfully tested and implemented, such as the DEEP methodology (Spector & Koszalka, 2004), KU-Mapper (Taricani & Clariana, 2006), and knowledge mapping tools (Herl, O’Neil, Chung, & Schacter, 1999; O’Neil, Chuang, & Baker, 2010). However, only a few of these instruments have been fully automated and tested for reliability and validity. Furthermore, automated and language-oriented assessment methodologies that enable a domain-independent analy-

---

This article was published Online First February 17, 2014.

Correspondence concerning this article should be addressed to Dirk Ifenthaler, Centre for Research in Digital Learning, Deakin University, Level 4, 550 Bourke Street, Melbourne VIC 3000, Australia. E-mail: dirk@ifenthaler.info



sis without a reference to large text corpora are scarce (Clariana, 2010).

There were three aims in the present study: (a) to introduce a language-oriented approach toward automated computer-based assessment and visualization of team-based performance that can be applied in educational large-scale assessments; (b) to investigate the feasibility and validity of the automated computer-based assessment and visualization methodology focusing on team performance as a specific cross-curricular skill, (c) suggesting a framework for integrating research and methodology development for the implementation of innovative computer-based assessment environments for international large-scale assessments (e.g., PISA, PIAAC).

### Team-Based Performance

A successful team typically possesses an informational advantage over individuals (Mesmer-Magnus & Dechurch, 2009). However, not all teams are able to take full advantage of these benefits. Some teams may even fail despite this advantage. Hence, there have been numerous attempts to identify the specific factors that make a team successful (Cannon-Bowers, Salas, & Converse, 1993; Guzzo & Dickson, 1996; Katzenbach & Smith, 1993; Sikorski, Johnson, & Ruscher, 2012; Van den Bossche, Gijssels, Segers, Woltjer, & Kirschner, 2011). For example, empirical research shows that through the use of combined resources, teams can successfully handle tasks and problems that otherwise would be too complex for a single individual (Badke-Schaub, Neumann, & Lauche, 2011; Bierhals, Schuster, Kohler, & Badke-Schaub, 2007; Cannon-Bowers & Salas, 2001; Cooke, Salas, Kiekel, & Bell, 2004; Eccles & Tenenbaum, 2004; Salas, Cooke, & Rosen, 2008).

Overall, shared mental models are regarded as a significant factor for successful team performance (Bandura, 1977, 1986; Cannon-Bowers et al., 1993; Cooke, Salas, Cannon-Bowers, & Stout, 2000; Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers, 2000; Van den Bossche et al., 2011). However, the concept of shared mental model is used and interpreted differently by various disciplines, for example, industrial/organizational psychology, human factors, social psychology, or system dynamics (Cooke et al., 2004). Within cognitive and educational psychology, the term *shared mental model* (SMM) is based on the theory of mental models (Johnson-Laird, 1983) and reflects internal representations that individuals construct to make sense of experiences with the world (Wittgenstein, 1922). Hence, individuals construct mental models in order to understand and explain experiences and events, process information, and solve complex problems (Gentner & Stevens, 1983; Johnson-Laird, 1989). More precisely, the theory of mental models is based on the assumption that cognitive processing takes place in the use of mental representations in which individuals organize symbols or representations of experience or thought in such a way that they effect a systematic representation of this experience or thought as a means of understanding or explaining it to others (Johnson-Laird, 1983). Hence, in order to create subjective plausibility, individuals construct an internal model that both integrates the relevant semantic knowledge and meets the perceived requirements of the situation (Ifenthaler & Seel, 2013). This internal model is referred to as an *individual mental model* (IMM). An SMM is denoted as a shared represen-

tation of a team that includes overlapping domain and task knowledge, skills, attitudes, objectives, processes, components, communication, coordination, adaption roles, relationships, behavior patterns, and interactions (Bandura, 1986; Cooke et al., 2004; Klimoski & Mohammed, 1994; Mohammed & Dumville, 2001).

Previous research shows that if team members share similar IMMs, they are more effective in their teamwork and perform better (Burke, Fiore, & Salas, 2003; Cannon-Bowers & Salas, 2001; Marks, Zaccaro, & Mathieu, 2000; Salas et al., 1992; Van den Bossche et al., 2011). For example, Lim and Klein (2006) found that shared task knowledge and shared team knowledge were valid predictors for team performance. Similar results regarding the influence of shared task and team knowledge on team performance were found in a series of studies using flight simulators in laboratory settings (Mathieu, Heffner, Goodwin, Cannon-Bowers, & Salas, 2005; Mathieu et al., 2000). Findings of a meta-analysis performed by Salas et al. (2008) suggest that team processes and cognitive as well as affective dispositions moderate performance outcomes of teams.

Although previous research highlighted different operationalizations of SMMs (Akkerman et al., 2007; Cooke et al., 2004), this study is based on an extended cognitive perspective of SMM. Figure 1 illustrates the interaction of IMM and SMM as well as its influence on team processes and team performance. The IMM of each team member integrates complex knowledge structures on declarative, procedural, causal, and metacognitive levels (Jonassen, Beissner, & Yacci, 1993; Kant, 1781/1998). The overlap of the IMMs constitutes the SMM. Cannon-Bowers and Salas (2001) identified two major components of an SMM: task-related components and team-related components. As every team member shares a certain amount of those components, it is therefore possible for a team to develop a collective understanding of tasks, conditions, and requirements that are needed to cope with a problem to be solved. However, this overlap is a result of complex interrelations between individual declarative, procedural, causal, and metacognitive knowledge as well as shared task and team-related knowledge (Cannon-Bowers & Salas, 2001). Team processes describe the transformation of all inputs through social interaction among team members into results, such as critical perspectives, new ideas, conflicts, decisions, or material objects. Finally, the result of all actions reflects the team performance (Bierhals et al., 2007; Mathieu et al., 2000).

### Automated Computer-Based Visualization and Assessment Methodology

Clearly, the direct assessment of an IMM or SMM is not possible (Jonassen & Cho, 2008; Miyake, 1986). Therefore, two classes of functions that describe the complex processes and interrelations between internal and external representations of mental models need to be considered (Ifenthaler, 2010c): (a)  $f_{in}$  as the function for the internal representation of objects of the world (internalization) and (b)  $f_{out}$  as the function for the external rerepresentation back to the world (externalization). None of these functions are directly observable either (Strasser, 2010). Thus, the assessment of IMMs and SMMs requires precise theoretical understanding and in-depth empirical investigations. Moreover, the possibilities of externalization for valid assessments are limited to

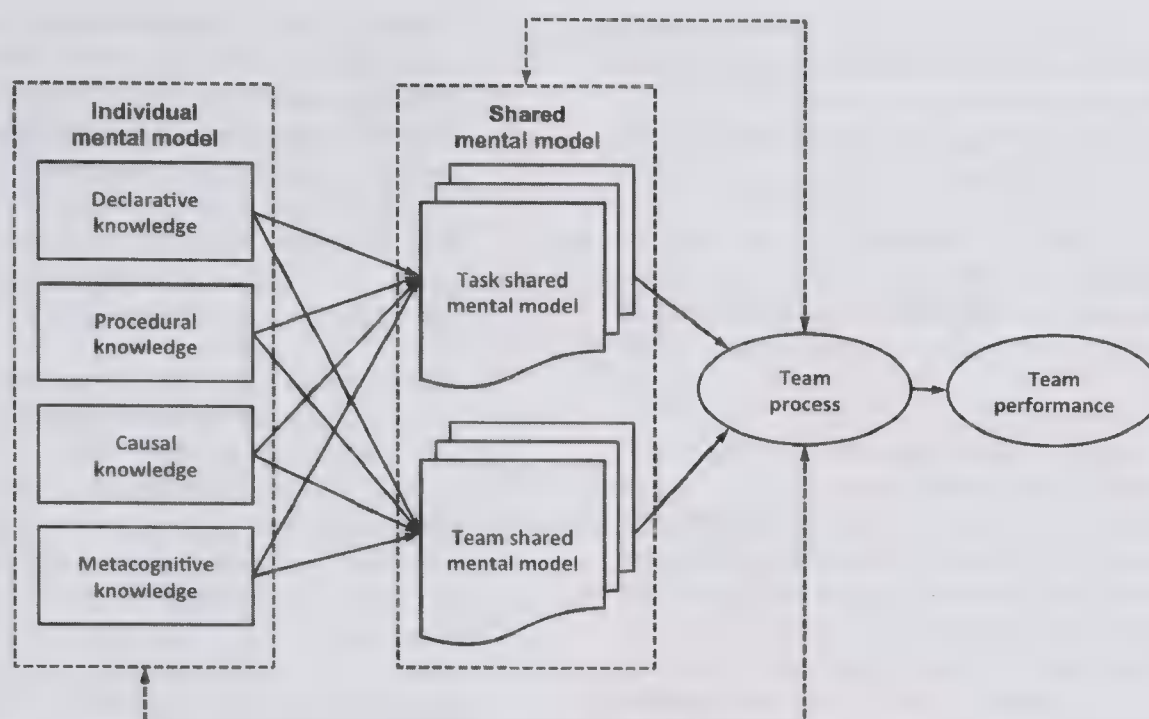


Figure 1. Interaction of individual mental model and shared mental model and its influence on team processes and team performance.

a few sets of sign and symbol systems, such as graphical or language-based approaches (Seel, 1999). Further, Minsky (1981) emphasized that different types of knowledge require different types of representations. Helbig (2006) argued that individuals are able to use different forms of representation of memorized information. They can either recall an appropriate form of representation from memory or transform memorized information in an appropriate form of representation in dependence on situational demands (Markman, 1999).

Because it is not possible to assess directly the internal representations, it is necessary to identify economic, fast, reliable, and valid methodologies to elicit and analyze externalized representations of IMMs and SMMs (Johnson, Ifenthaler, Pirnay-Dummer, & Spector, 2009). Despite recent advances in educational technology, empirical research and application of *fully* automated assessment methodologies in complex domains (Greiff, Wüstenberg, Molnár, et al., 2013) and IMMs and SMMs are very limited (Carley, 1997; Fischer, Bruhn, Gräsel, & Mandl, 2002; Fischer & Mandl, 2005; Mohammed, Klimoski, & Rentsch, 2000).

The newly developed and fully automated Automated Knowledge Visualization and Assessment (AKOVIA) methodology is based on mental model theory (Johnson-Laird, 1989) and integrates a large number of dynamic interfaces to different online environments, for instance, learning management systems, personalized learning environments, game-based environments, or computer-based assessment environments such as PISA or PIAAC. This open architecture of AKOVIA enables a large variety of research and practical applications, such as investigation of learning processes; distinguishing features of subject domains; cross-curricular, nonroutine, dynamic, and complex skill; or convergence of team-based knowledge.

### Analysis Algorithm

The underlying assumption of AKOVIA is that IMMs and SMMs can be externalized and rerepresented as a graph (Rumelhart & Norman, 1978). A graph consists of a set of vertices whose relations are represented by a set of edges (Tittmann, 2010). Various measures from graph theory have been successfully applied in previous studies in order to analyze externalized IMMs and SMMs and their development over time within different environments and domains (Ifenthaler, Masduki, & Seel, 2011; O'Neil et al., 2010; Schvaneveldt, 1990). However, most of these studies used graphical rerepresentations (e.g., knowledge map, causal map) to assess the externalized IMM or SMM (Chung et al., 1999).

Using a natural language-oriented approach limits the bias of externalization (e.g., through causal maps, which require extensive training), as language is regarded as the most automated and direct form of externalization (Chomsky, 1970; Searle & Grewendorf, 2002). Given the latest developments in educational technology, language-based approaches have been developed, such as text classification and machine learning methodologies, which allow for the automatic processing and analyzing of texts in various forms (Gweon, Rosé, Wittwer, & Nueckles, 2005).

AKOVIA integrates such a language-oriented methodology. It follows the axiom on association and sequences: What is closely related is also closely externalized (Frazier, 1999; Pollio, 1966). The methodology relies on the dependence of syntax and semantics within natural language and uses the associative features of text as a heuristic to rerepresent knowledge. Unlike approaches from latent semantic analysis (Foltz, Kintsch, & Landauer, 1998), Web ontologies and semantic Web (Ding, 2001), the language-oriented approach of AKOVIA can operate on a comparably small amount of text (approximately 300 words) and does not require a



reference to large (web-based) text corpora. In order to provide cross-curricular applications, AKOVIA is operating domain independently on available text input into the system.

AKOVIA's language-oriented analysis is carried out in multiple stages (Ifenthaler & Pirnay-Dummer, 2014):

1. Text input and cleaning is where the text is taken into the system (upload function through a web browser or via database interface) and checked for a specified character set, including the deletion of metadata such as HTML tags.
2. Text parsing, stemming, and calculation of word associations is where the text is split into sentences and single tokens such as words, punctuation marks, and quotation marks. Through a rule-based and corpus-based tagging process, nouns and names are identified within the text (Brill, 1995). Next, the stemming process reduces all words to their word stems as different inflections of a word need to be treated as one in the further analysis process (e.g., singular and plural forms such as *book* and *books*). Then, the associatedness is calculated by (a) identifying the default length of sentences, that is, the longest sentence in the text plus one and counting the number of words for each individual sentence; (b) identifying all possible pairs of words; (c) calculating the distance between words of all pairs within all sentences, that is, the minimum number of words between the words of the pair in a single sentence; (d) calculating the sum of distances for the text for all pairs; (e) building a hierarchy of distances for all pairs; (f) final output generation including the lowest sum of distances, that is, only pairs with association evidence in the text are included.

3. The graph-based analyses include seven measures that enable a quantitative description of structural and semantic features of the text (see Table 1). The structural and semantic comparisons identify similarities of frequencies or sets of properties between texts, for example, expert text versus novice text (Goldsmith & Davenport, 1990). The quantitative measures (see Table 1) are defined between  $s = 0$  (complete exclusion) and  $s = 1$  (complete identity);  $0 \leq s \leq 1$  (Tversky, 1977).
4. The graphical output is realized with the help of the open-source graph visualization software GraphViz (Ellson, Gansner, Koutsofios, North, & Woodhull, 2003). GraphViz uses the list form of the hierarchy of distances for all pairs and constructs a vertex-edge-vertex representation of the most frequent pairs (see Figure 2). Each vertex of the graphical output contains a destemmed word of the pairs. The edge of the graphical output contains an indicator for the noun-distance generated from the text (Pirnay-Dummer & Ifenthaler, 2011). Additionally, different colors of the edges indicate the strength of association. The graphical output can also be generated without the indicators on the edges.

The automated language-oriented analysis can be applied domain independently for written texts (e.g., essay text) or graphical representations (e.g., causal map, concept map) against a single or multiple reference models (Coronges, Stacy, & Valente, 2007). The reference model can be either the same individual's prior understanding of a phenomenon in question, another team member's understanding, a shared or aggregated understanding of the phenomenon, or an expert solution of the phenomenon in question. Cross-comparisons between written texts and graphical represen-

Table 1  
*Description of the Seven AKOVIA Measures*

Measure [abbreviation] and type	Short description
Surface matching [SFM] <i>Structural indicator</i>	The surface matching compares the number of vertices within two graphs. It is a simple and easy way to calculate values for surface complexity.
Graphical matching [GRM] <i>Structural indicator</i>	The graphical matching compares the diameters of the spanning trees of the graphs, which is an indicator for the range of conceptual knowledge. It corresponds to structural matching as it is also a measure for structural complexity only.
Structural matching [STM] <i>Structural indicator</i>	The structural matching compares the complete structures of two graphs without regard to their content. This measure is necessary for all hypotheses that make assumptions about general features of structure (e.g., assumptions, which state that expert knowledge is structured differently from novice knowledge).
Gamma matching [GAM] <i>Structural indicator</i>	The gamma matching describes the quotient of terms per vertex within a graph. Because both graphs that connect every term with each other term (everything with everything) and graphs that only connect pairs of terms can be considered weak models, a medium density is expected for most good working models.
Concept matching [CCM] <i>Semantic indicator</i>	Concept matching compares the sets of concepts within a graph to determine the use of terms. This measure is especially important for different groups that operate in the same domain (e.g., use the same textbook). It determines differences in language use between the models.
Propositional matching [PPM] <i>Semantic indicator</i>	The propositional matching value compares only fully identical propositions between two graphs. It is a good measure for quantifying semantic similarity between two graphs.
Balanced semantic matching [BSM] <i>Semantic indicator</i>	The balanced semantic matching is the quotient of propositional matching and concept matching. In specific cases (e.g., when focusing on complex causal relations), balanced propositional matching could be preferred over propositional matching.

Note. AKOVIA = Automated Knowledge Visualization and Assessment.

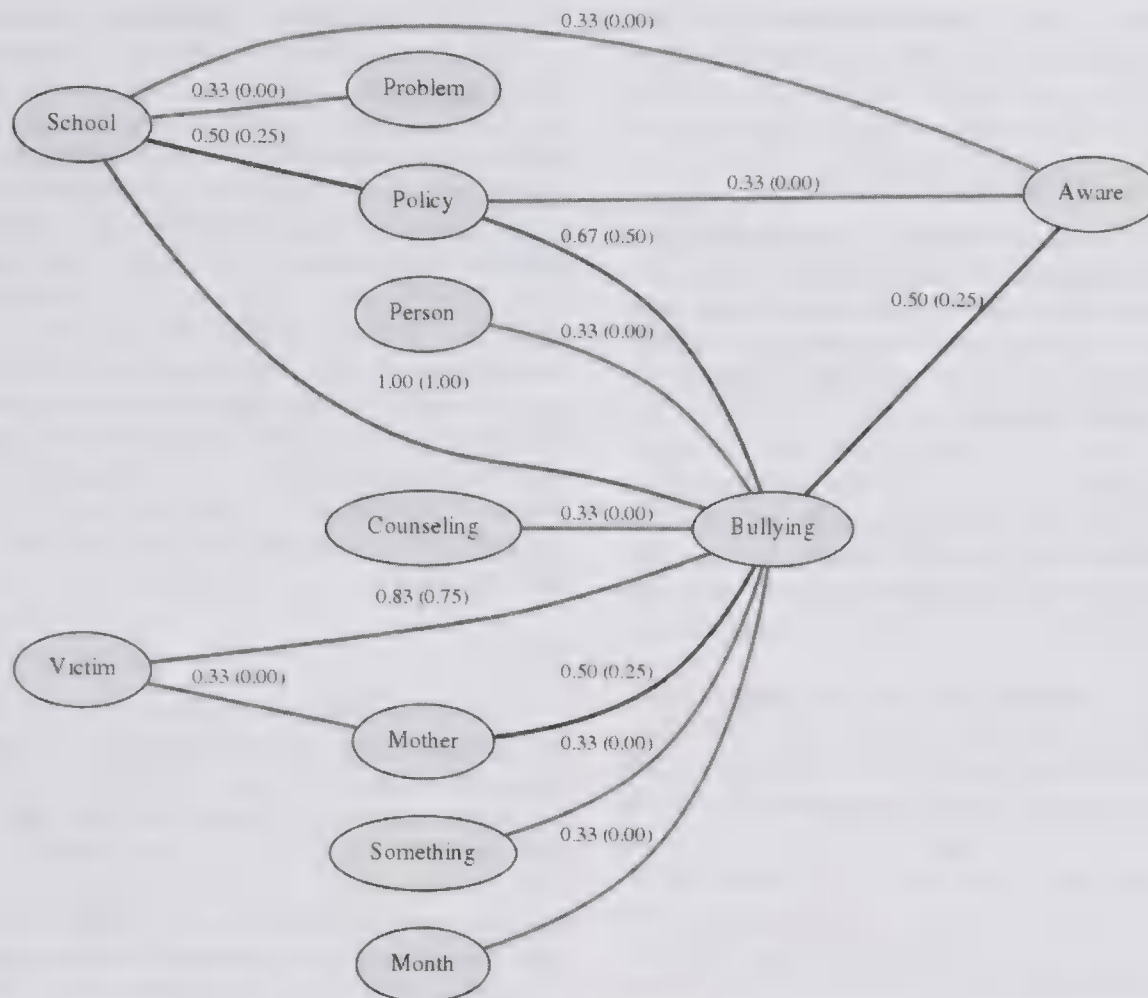


Figure 2. Automated Knowledge Visualization and Assessment standardized graphical output. The association strength between concepts is displayed by the numbers on the links. The value outside the brackets shows the weight from the list form; the second value inside the bracket displays the weight relative to what is actually visualized.

tations are also possible (Ifenthaler, 2011). For team-based assessment, the aggregate function allows the grouping of individual representations into an aggregated representation (Pirnay-Dummer & Ifenthaler, 2010).

### Test Quality and Application

The underlying analysis algorithms of AKOVIA have been successfully tested for reliability and validity in various experimental studies focusing on IMMs (Al-Diban & Ifenthaler, 2011; Johnson et al., 2011; McKeown, 2009; Pirnay-Dummer, Ifenthaler, & Spector, 2010). Reliability scores exist for single measures integrated into AKOVIA. They range from  $r = .79$  to  $.94$  and are tested for the structural and semantic measures separately and across different domains. Validity scores are also reported separately for the structural and semantic measures (convergent validity  $r = .71$  to  $r = .91$ ) (Pirnay-Dummer & Ifenthaler, 2010).

Kim (2012) as well as Al-Diban and Ifenthaler (2011) conducted cross-validation studies in order to identify the test quality of the underlying AKOVIA analysis algorithms. Both studies identified acceptable robustness of the automatically generated results when compared with traditional manual analysis procedures such as qualitative content analysis.

In a recent longitudinal study using the AKOVIA methodology, an in-depth hierarchical linear modeling analysis revealed patterns

of the learning-dependent progression of IMMs on structural and semantic levels (Ifenthaler et al., 2011). In a series of experimental studies, the effectiveness of preflective (Ifenthaler & Lehmann, 2012) and reflective (Ifenthaler, 2012) prompts for self-regulated learning within problem-solving processes were successfully investigated with the AKOVIA methodology. Another experimental study compared domain-distinguishing features of IMMs using the structural and semantic comparison functions of AKOVIA (Ifenthaler, 2011). The results showed unique features of the biology, history, and mathematics domains. The AKOVIA methodology was also applied in order to compare unique features of written essays and causal representations (Johnson et al., 2011). Johnson et al. (2011) identified in their study significant differences of structural properties and semantic content of written texts and causal representations when produced by the same learner within identical domains. Pirnay-Dummer and Ifenthaler (2011) applied the AKOVIA methodology for automatically generating feedback models on the fly. They found that the automatically generated feedback models had identical impact on problem solving when compared with feedback models generated by domain experts. Another series of experimental studies investigated variations of automated feedback models (standardized graphical output) generated with AKOVIA (Ifenthaler, 2009, 2010a). These studies highlight the benefits of automated feedback models for online



learning environments and self-regulated learning because they can offer scaffolding and feedback whenever the learner requires it.

Research Questions and Hypotheses

The central research objective was to advance knowledge in automated and language-oriented computer-based assessment of team-based performance that can be applied in educational large-scale assessments. Therefore, this study was designed to investigate the feasibility and validity of the AKOVIA assessment and visualization methodology focusing on team performance as a specific cross-curricular skill. Specifically, we tested (a) the underlying analysis algorithm assuming that the structural and semantic measures of AKOVIA precisely identify similarities and differences in team-based performance. Further, we tested (b) whether the team-related knowledge was related to team-based performance identified through the structural and semantic measures of AKOVIA.

With regard to (a), past investigations link team-based performance to individual knowledge and abilities as well as identified that a collective lack of knowledge, skills, and resources lead to ineffective performance (Kozlowski & Ilgen, 2006; Moreland & Levine, 1992). Hence, numerous concepts have been postulated in order to test the composition of individual knowledge and characteristics on team performance (Chambers & Abrami, 1991; Marks et al., 2000). Among those, task-related knowledge of individual team members has been identified as a critical predictor for team-based performance (Horwitz, 2005), suggesting that diversity among team members as well as high knowledge levels have a positive effect on the team’s performance (Cox & Blake, 1991; Hambrick, Cho, & Chen, 1996). As part of assessing the validity of the structural and semantic measures of AKOVIA, we adhere to the question whether teams composed of different levels of task-related knowledge will show superior performance relative to homogeneously composed teams.

*Hypothesis 1:* It is hypothesized that the structural (Hypothesis 1a) and semantic (Hypothesis 1b) measures of AKOVIA provide evidence for differences of team performance between differently composed teams based on their task-related knowledge.

With regard to (b), it is commonly accepted that team-related knowledge predicts team-based performance (Badke-Schaub et al., 2011; Cannon-Bowers & Salas, 2001; Johnson & Lee, 2008). For example, Lim and Klein (2006) identified that team-related knowledge influenced team performance. More precisely, their findings suggest that task shared mental models (TaSMMs) and team shared mental models (TeSMMs) were positively predicting team performance. Similar findings have been reported by Mathieu et al. (2000). Accordingly, the relation between SMMs and team-based performance will be further indicators of validity for the structural and semantic measures of AKOVIA.

*Hypothesis 2:* It is hypothesized that greater levels of TaSMMs (Hypothesis 2a) and TeSMMs (Hypothesis 2b) are associated with higher team-based performance assessed with the structural and semantic measures of AKOVIA.

Method

Participants

Participants were freshman university students enrolled in an introductory psychology course at a German university. They participated for extra course credit. Five students who did not provide complete data were excluded from the sample. The final sample consisted of 224 students (68 men, 156 women; mean age = 21.6 years, *SD* = 2.48, min = 18, max = 33). They studied an average of 2.83 semesters (*SD* = 2.92). For 90% of the participants, it was their first enrollment in a university program, and 19% of the participants reported that they successfully finished a formal vocational training before the current university education. Of the participants, 79% rated their ability to work in a team as high; 81% rated their computer and social media skills as medium to high.

Design

A three-step computer-based algorithm was used to assign participants to teams and experimental conditions. The first step determined the participant’s total score of the introductory domain-specific knowledge test and the verbal abilities test (see the Measures section for detailed description of instruments). The second step determined the range of total scores for the three experimental conditions: (a) low knowledge team (LKT) composed of individual participants with low total scores (domain-specific knowledge and verbal abilities tests), (b) high knowledge team (HKT) composed of individual participants with high total scores, and (c) mixed knowledge team (MKT) composed of individual participants with low and high total scores (see Table 2; see also the Measures section for a detailed description of instruments). The third step randomly assigned participants to teams of four members, resulting in 56 teams (LKT, *n*<sub>1</sub> = 18 teams; HKT, *n*<sub>2</sub> = 18 teams; MKT, *n*<sub>3</sub> = 20 teams). Tasks, materials, and procedure were identical for all experimental conditions.

Online Environment and Team Tasks

The online environment included all necessary information (tasks, reading materials, measures, feedback, contacts) for the individual participants and teams. Participants used a self-generated unique identifier to gain access to the online environment. After the first access to the online environments, participants were prompted to contact their team members. The study included two tasks (Ta1, Ta2)

Table 2  
*Means and Standard Deviations of Domain-Specific Knowledge Test and the Verbal Abilities Test*

Variable	LKT		HKT		MKT	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Domain-specific knowledge pretest	2.86	1.04	5.10	1.35	3.79	1.27
Domain-specific knowledge posttest	4.38	1.81	6.25	1.42	5.23	1.34
Verbal abilities test	5.22	1.26	8.92	2.13	6.95	2.01

*Note.* LKT = low knowledge team; HKT = high knowledge team; MKT = mixed knowledge team.

to be solved by the teams in the form of a written essay (see Table 3 for a detailed description of tasks and instructions). For each task, a domain expert generated a reference solution on the basis of the available learning materials in the form of a written essay. The tasks were generated automatically as a portable document format (PDF) document and were available for download any time participants accessed the online environment. The document included the participant's unique identifier, team number, and team members as well as all task-related instructions including references to the learning materials (see Table 3). An upload function in the online environment was available for handing in the task solution within a predefined time frame (72 hr after availability of the task). All team members had to upload the agreed-upon team's solution in order to control for their participation and to prompt them to additional assessments. Accordingly, at specific points of the study, participants were asked to complete short questionnaires in the online environment (see the Procedure section for details about the sequence and frequency of the applied assessments).

## Measures

**Domain-specific knowledge.** The domain-specific knowledge test included 11 multiple-choice questions with five possible solutions each (one correct, four incorrect). They were developed on the basis of introductory reading materials available to all students of the psychology course with a special focus on learning and assessment. A pilot study ( $N = 8$  participants, independent from the participants of the main study) was used to test the average difficulty level to account

for ceiling effects. Two identical versions (in which the 11 multiple-choice questions appeared in a different order) of the domain-specific knowledge tests (pre- and posttest) were administered. For example, items administered included "What is the definition of classical conditioning?" or "What does the forgetting curve hypothesize?" It took about 6 min to complete the test.

**Verbal abilities.** Participants' verbal abilities were tested with a subscale of the I-S-T 2000R intelligence test (Intelligenz-Struktur-Test; Amthauer, Brocke, Liepmann, & Beauducel, 2001). A total of 20 sentences with a missing word had to be completed using a set of five words. Overall, the widely used intelligence test has a high reliability ( $r = .88-.96$ ; split-half reliability). It took about 6 min to complete the test.

**Team Assessment Diagnostic Measure (TADM).** The TADM questionnaire (Johnson, Lee, Lee, & O'Connor, 2007) identifies two team-related factors: (a) the TaSMM (eight items) and (b) the TeSMM (seven items). The 15 items were answered on a 5-point Likert scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *not sure*, 4 = *agree*, 5 = *strongly agree*). For example, an item related to the TeSMM reads: "My team communicates effectively with each other while performing our tasks." Johnson et al. (2007) reported acceptable content validity and successful factorial structure of the instrument. Cronbach's alpha ranges from .75 to .89.

## Procedure

In the first phase of the study, the participants created a unique identifier when accessing the online environment and completed a

Table 3  
Tasks Available in the Online Environment as Downloadable PDF Document (Translated From German)

Measurement point	Task description and instruction
Task 1	<p>Please work with your team members to respond to the following question in the form of a written text (one page), which shall be published in a teacher magazine: <i>How did the concept of learning change during the 20th century?</i></p> <p>[Link to key reference material]</p> <p>Contact your three team members (see contact details below) and organize at least one synchronous virtual meeting (e.g., via Skype) in which you discuss and solve your task. Each of your team members will need to upload the solution in the online environment within the next 72 hours. After uploading the solution, you will be prompted to answer two short questionnaires.</p> <p>[Personal identifier] [Team number] [Names and contact information of team members] [Link to online environment] [Contact information of examiner]</p>
Task 2	<p>Please work with your team members to respond to the following question in the form of a written text (one page), which shall be published in a teacher magazine: <i>Which functions do performance assessments in schools have?</i></p> <p>[Link to key reference material]</p> <p>Contact your three team members (see contact details below) and organize at least one synchronous virtual meeting (e.g., via Skype) in which you discuss and solve your task. Each of your team members will need to upload the solution in the online environment within the next 72 hours. After uploading the solution, you will be prompted to answer two short questionnaires.</p> <p>[Personal identifier] [Team number] [Names and contact information of team members] [Link to online environment] [Contact information of examiner]</p>



demographic data questionnaire as well as the introductory domain-specific knowledge test and verbal abilities test. As outlined above (see the Design section), the total score of the introductory tests were automatically generated, and teams (four members each) were randomly composed on the basis of the results of the tests and the definition of experimental conditions (LKT, HKT, MKT). After 72 hr, participants accessed the online environment and were prompted to the first task (Ta1), including a detailed instruction and the contact information of the team members (all as a downloadable PDF document) (see Table 3). The teams were asked to meet at least once for a synchronous virtual interaction (e.g., Skype) within the next 72 hr. Each team member completed the team assessment and diagnostic measure (TADM) before uploading the team solution within the allocated 72 hr. In the third phase of the study, participants accessed the online environment and were promoted to the second task (Ta2; a downloadable PDF document including instructions and contact information of the team members) (see Table 3). After solving the second task in the team, including at least one synchronous virtual interaction (e.g., via Skype) within 72 hr, each team member completed the TADM questionnaire before uploading the team solution within the allocated 72 hr. Finally, students completed the postversion of the domain-specific knowledge test.

### Variables and Data Analyses

The quality of the team-based performance (written essays) was analyzed with AKOVIA by comparing them against the reference solution, which was based on an expert solution of the individual task (Ta1, Ta2) and the available learning materials. This study uses two AKOVIA measures: Gamma Matching (GAM) and Balanced Semantic Matching (BSM). BSM, as the quotient of all unique semantic AKOVIA measures, has proved preferable for semantic comparisons of written texts as it includes the semantic information of single concepts *and* more complex semantic information of propositions (Johnson et al., 2011; Pirnay-Dummer & Ifenthaler, 2011). Other studies successfully used GAM for identifying the complexity within several subject domains as it reflects the structural connectedness of knowledge externalizations (Ifenthaler et al., 2011; McKeown, 2009). Reliability coefficients of the administered instruments are acceptable and consistent with previously reported results: verbal abilities (split-half reliability = .982), TaSMM<sub>1</sub> (Cronbach's  $\alpha$  = .840), TaSMM<sub>2</sub> (Cronbach's  $\alpha$  =

.876), TeSMM<sub>1</sub> (Cronbach's  $\alpha$  = .835), TeSMM<sub>2</sub> (Cronbach's  $\alpha$  = .851). Table 4 shows all variables and scoring specification of the study.

We conducted an analysis of variance (ANOVA) to analyze between-group differences by experimental groups (Hypothesis 1). In order to control for Type I error, Tukey's honestly significant difference (HSD) post hoc comparisons were used to examine differences between experimental groups. As a second major analytic strategy, regression models were performed, one for each of the four team-based performance measures as the dependent variables (Hypothesis 2): GAM<sub>t</sub> (structural measure) and BSM<sub>t</sub> (semantic measure).

## Results

### Descriptive Analyses

The average text length of the responses for Ta1 was  $M = 379.79$  words ( $SD = 85.43$ ; min = 238; max = 546) and for Ta2  $M = 385.55$  words ( $SD = 82.27$ ; min = 275; max = 546). Accordingly, the required text length for a valid AKOVIA analysis (approximately 300 words) was met. No significant differences of text length were found between the experimental groups (LKT, HKT, MKT) for Ta1,  $F(2, 221) = 0.375$ , *ns*, and Ta2,  $F(2, 221) = 0.137$ , *ns*, as well as between the two tasks,  $t(223) = 1.254$ , *ns*.

On the domain-specific knowledge test (pre- and posttest), participants could score a maximum of 11 correct answers. In the pretest (DKpre), they scored an average of  $M = 3.91$  correct answers ( $SD = 1.52$ ; min = 0; max = 8), and in the posttest (DKpost) they scored an average of  $M = 5.28$  correct answers ( $SD = 1.70$ ; min = 0; max = 10). The increase in correct answers was significant,  $t(223) = 13.863$ ,  $p < .001$ ,  $d = 1.857$ .

### Hypothesis 1: Differently Composed Teams

Table 5 summarizes the means and standard deviations for the structural (GAM) and semantic (BSM) team-based performance for each task (Ta1, Ta2) and the three experimental groups (LKT, HKT, MKT).

**Ta1.** With regard to Hypothesis 1a (structural AKOVIA measure indicating differences in team performance), an ANOVA revealed significant differences between the three experimental groups for the structural team-based performance measure (GAM<sub>1</sub>),  $F(2, 221) = 48.037$ ,  $p < .001$ ,  $\eta^2 = .303$ . Tukey's HSD

Table 4  
Variables of the Study, Corresponding Instrument, and Description of Scoring

Variable [abbreviation]	Instrument	Scoring
Gamma matching [GAM <sub>t</sub> ]	AKOVIA	AKOVIA structural similarity measure; $0 \leq GAM_t \leq 1$ ; for task; $t = 1, 2$
Balanced semantic matching [BSM <sub>t</sub> ]	AKOVIA	AKOVIA semantic similarity measure; $0 \leq BSM_t \leq 1$ ; for task; $t = 1, 2$
Domain-specific knowledge (pretest) [DKpre]	Multiple-choice test	Sum of correct answers; $0 \leq DKpre \leq 11$
Domain-specific knowledge (posttest) [DKpost]	Multiple-choice test	Sum of correct answers; $0 \leq DKpost \leq 11$
Domain-specific knowledge gain [DKgain]	Multiple-choice test	$DKgain = DKpost - DKpre$ ; $-11 \leq DKgain \leq 11$
Verbal abilities [VA]	I-S-T 2000R	Sum of correct answers; $0 \leq VA \leq 11$
Task shared mental model [TaSMM <sub>t</sub> ]	TADM	Mean rating of scale items; $1 \leq TaSMM_t \leq 5$ ; for task; $t = 1, 2$
Team shared mental model [TeSMM <sub>t</sub> ]	TADM	Mean rating of scale items; $1 \leq TeSMM_t \leq 5$ ; for task; $t = 1, 2$

Note. AKOVIA = Automated Knowledge Visualization and Assessment; I-S-T 2000R = Intelligenz-Struktur-Test; TADM = Team Assessment Diagnostic Measure.

Table 5  
Means and Standard Deviations of Team-Based Performance

Task	LKT		HKT		MKT	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
GAM						
Ta1	.646	.145	.802	.147	.843	.091
Ta2	.674	.181	.684	.172	.647	.187
BSM						
Ta1	.502	.254	.622	.129	.748	.036
Ta2	.605	.176	.764	.133	.826	.094

Note. LKT = low knowledge team; HKT = high knowledge team; MKT = mixed knowledge team; Ta1 = Task 1; Ta2 = Task 2; GAM = gamma matching; BSM = balanced semantic matching.

post hoc comparisons indicate that the HKTs (95% CI [.767, .836],  $p < .001$ ) and MKTs (95% CI [.823, .863],  $p < .001$ ) gained significantly higher structural similarity than the LKTs (95% CI [.612, .680]). However, comparisons between the HKTs and MKTs were not statistically significant at  $p < .05$ .

A second ANOVA was computed in order to test Hypothesis 1b (semantic AKOVIA measure indicating differences in team performance). An ANOVA revealed a significant difference between the three experimental groups for the semantic team-based performance measure (BSM<sub>1</sub>),  $F(2, 221) = 43.351$ ,  $p < .001$ ,  $\eta^2 = .282$ . Tukey's HSD post hoc comparisons indicate that the HKTs (95% CI [.592, .652],  $p < .001$ ) and MKTs (95% CI [.740, .756],  $p < .001$ ) gained significantly higher semantic similarity than the LKTs (95% CI [.442, .562]). The Tukey's HSD comparisons between the HKTs and MKTs was statistically significant at  $p < .001$ .

**Ta2.** With regard to Hypothesis 1a (structural AKOVIA measure indicating differences in team performance), no significant between-group differences could be identified across the three experimental groups for the structural team-based performance measure (GAM<sub>2</sub>),  $F(2, 221) = 0.831$ ,  $ns$ .

An ANOVA was computed in order to test Hypothesis 1b (semantic AKOVIA measure indicating differences in team performance). A significant difference was found between the three experimental groups for the semantic team-based performance (BSM<sub>2</sub>),  $F(2, 221) = 51.559$ ,  $p < .001$ ,  $\eta^2 = .318$ . Tukey's HSD post hoc comparisons indicate that the HKTs (95% CI [.732, .795],  $p < .001$ ) and MKTs (95% CI [.805, .847],  $p < .001$ ) gained significantly higher semantic similarity than the LKTs (95% CI [.563, .646]). The Tukey's HSD comparisons between the HKTs and MKTs was statistically significant at  $p = .016$ .

To sum up, the expected differences of team-based performance between differently composed teams, based on their task-related knowledge, were found on both the structural (GAM) and semantic (BSM) measures of AKOVIA.

## Hypothesis 2: Influence of SMMs

For each outcome variable (structural team performance, GAM<sub>1</sub>; semantic team performance, BSM<sub>1</sub>) and task (Ta1, Ta2), we conducted a regression analysis accounting for prediction by TaSMM<sub>1</sub> and TeSMM<sub>1</sub>.

**Ta1.** Table 6 shows the zero-order correlations of predictors used in the regression analyses for Ta1, indicating significant correlations between structural/semantic team performance and Ta/TeSMM.

The results of the regression analyses for structural (GAM<sub>1</sub>) and semantic (BSM<sub>1</sub>) team-based performance are presented in Table 7, yielding a  $\Delta R^2$  of .163 and .244 for GAM<sub>1</sub> and BSM<sub>1</sub>, respectively. For GAM<sub>1</sub>, the SMM contributed unique variance to the structural team-based performance. Specifically, TaSMM<sub>1</sub> and TeSMM<sub>1</sub> positively predicted the structural team-based performance (GAM<sub>1</sub>), indicating that the higher the sharedness of task and team knowledge, the higher the structural team-based performance (see Table 7). For BSM<sub>1</sub>, the SMM contributed unique variance to the semantic team performance. Specifically, TaSMM<sub>1</sub> and TeSMM<sub>1</sub> positively predicted the semantic team-based performance (BSM<sub>1</sub>), indicating that the higher the sharedness of task and team knowledge, the higher the semantic team-based performance (see Table 7).

**Ta2.** Table 8 shows the zero-order correlations of predictors used in the regression analyses for Ta2, indicating significant correlations between semantic team performance and TeSMM.

The regression analysis for structural team-based performance (GAM<sub>2</sub>) did not explain a significant amount of variance ( $\Delta R^2 = .001$ ),  $F(2, 221) = 1.114$ ,  $ns$ . For the semantic team-based performance (BSM<sub>2</sub>), the regression model explained  $\Delta R^2 = .114$  of variance (see Table 7). Specifically, TaSMM<sub>2</sub> positively predicted the semantic team-based performance (BSM<sub>2</sub>), indicating that the higher the sharedness of task knowledge, the higher the semantic team-based performance (see Table 7).

To sum up, for Ta1, the findings suggest that higher structural and semantic team-based performance were predicted by greater levels of sharedness of task and team knowledge. For Ta2, the findings suggest that higher semantic team-based performance was predicted by greater levels of sharedness of task knowledge.

## Discussion

Cooke et al. (2000) reviewed the strengths and weaknesses of methods for assessing team knowledge, including observations, interviews, questionnaires, process tracing, and knowledge mapping. None of these methods, however, used automated algorithms for natural language-oriented assessment of team performance. Hence, in this study, the feasibility and validity of a language-oriented approach toward automated computer-based assessment and visualization of team-based performance was investigated,

Table 6  
Descriptives and Zero-Order Correlations of Predictor Variables for Task 1

Variable	1	2	3	4
1. Task shared mental model	—			
2. Team shared mental model	.321**	—		
3. Structural team performance	.381**	.274**	—	
4. Semantic team performance	.454**	.345**	.390**	—
<i>M</i>	3.87	3.95	.77	.63
<i>SD</i>	.67	.56	.15	.19

\*\*  $p < .01$ .



Table 7

Regression Analysis Predicting Structural ( $GAM_i$ ) and Semantic ( $BSM_i$ ) Team Performance for Task 1 and Task 2

Variable	Task 1									
	$GAM_1$					$BSM_1$				
	$R^2$	Adjusted $R^2$	$B$	$SE\ B$	$\beta$	$R^2$	Adjusted $R^2$	$B$	$SE\ B$	$\beta$
Shared mental model	.171	.163				.250	.244			
Task shared mental model			.076	.015	.327***			.110	.018	.383***
Team shared mental model			.046	.018	.169**			.076	.021	.222***
Variable	Task 2									
	$GAM_2$					$BSM_2$				
	$R^2$	Adjusted $R^2$	$B$	$SE\ B$	$\beta$	$R^2$	Adjusted $R^2$	$B$	$SE\ B$	$\beta$
Shared mental model	.010	.001				.122	.114			
Task shared mental model			-.013	.025	-.038			.116	.021	.374***
Team shared mental model			.033	.022	.109			-.022	.019	-.082

\*\*  $p < .01$ . \*\*\*  $p < .001$ .

which can be applied in educational large-scale assessments (e.g., PISA, PIAAC).

The computer-based AKOVIA methodology integrates a multi-stage language-oriented algorithm that transforms text into a list form and a proximity matrix by assigning distances and weights to single words and identifying associative evidence in sentences and the text (Pirnay-Dummer & Ifenthaler, 2010). The resulting list form and proximity matrix enables an in-depth analysis of structural and semantic features of the text. Currently, AKOVIA supports four structural, three semantic, and additional graph theory-based measures (see Table 1). The structural measures identify surface features of the knowledge representation such as the sum of concepts (SFM; Surface Matching), the complexity of concepts (GRM; Graphical Matching), and the connectedness of concepts (GAM). Additionally, deep structural features can be analyzed by deconstructing the knowledge representation into the smallest possible units (STM; Structural Matching). The semantic measures operate on the stemmed words of the knowledge representation identifying predefined semantic features of single words (CCM; Concept Matching) or propositions (PPM; Propositional Matching) defined as a word linked to another word (Ifenthaler, 2010b). The quotient of PPM and CCM results in the BSM. Graph theory-based measures provide evidence about the connectedness of the knowledge representation (i.e., all concepts are linked to reach every concept from every other concept). Ruggedness indicates the

sum of subrepresentations, that is, independent concepts or propositions not linked to other parts of the knowledge representation (Ifenthaler et al., 2011; Schvaneveldt, 1990). The standardized graphical output is constructed from the  $N$  strongest relations within the whole proximity matrix of the knowledge representation using GraphViz (Ellson et al., 2003).  $N$  can be set within the AKOVIA analysis functions in order to accommodate specific assessment and analysis requirements such as limited word length of available texts.

As suggested by previous empirical research (Horwitz & Horwitz, 2007), our findings revealed significant differences of structural and semantic team-based performance between differently composed teams, that is, low task-related knowledge, high task-related knowledge, and mixed task-related knowledge. More specifically, the results of this study suggest that AKOVIA's BSM is an acceptable measure for language-oriented assessment of team-based performance. BSM balances the dependency of semantically correct concepts (vertices) and causal relations (i.e., propositions) of semantically correct concepts as well as includes structural features (Ifenthaler, 2010c). Whereas CCM only identifies the semantically correct use of single concepts and PPM only identifies the use of semantically correct propositions, BSM accounts for all of these features. Additionally, the expected differences of team-based performance between differently composed teams, based on their task-related knowledge, were found for the structural (GAM) measure of AKOVIA.

Further, results suggest that higher structural and semantic team-based performance measured with AKOVIA were predicted by greater levels of sharedness of task and team knowledge. The identified relations between SMMs and team-based performance are consistent with previous research (Badke-Schaub et al., 2011; Cannon-Bowers & Salas, 2001). However, the results for the two different tasks were not fully consistent as TeSMM was not a significant predictor for GAM and BSM. This could be attributed to the TeSMM factor of the TADM questionnaire, which was not tested and implemented with repeated measures designs in previous studies. Accordingly, further empirical investigations with TADM should focus on the consistency of the instrument over

Table 8

Descriptives and Zero-Order Correlations of Predictor Variables for Task 2

Variable	1	2	3	4
1. Task shared mental model	—			
2. Team shared mental model	.406**	—		
3. Structural team performance	.006	.094	—	
4. Semantic team performance	.341**	.070	.060	—
$M$	4.04	3.93	.67	.73
$SD$	.53	.60	.18	.17

\*\*  $p < .01$ .

time. A further data analysis with additional structural AKOVIA measures (SFM, GRM, STM) confirmed the result of the GAM measure. No correlation between the TeSMM factor and structural AKOVIA measures were found.

Hence, the findings of this study provide initial but resilient evidence of the feasibility and validity of the AKOVIA methodology for an automated and language-oriented computer-based assessment of team performance. As the analysis algorithm is scalable and adaptive to educational settings and assessments, application within international large-scale assessments should be explored in the future. Still, limitations of this study need to be addressed, and further empirical research is required to replicate and advance the findings of this study.

## Limitations and Future Research

As with all experimental research, there are obvious limitations to this study that require consideration, primarily referring to sample characteristics and methodological issues.

First, the conceptual model focusing on the interaction of IMMs and SMMs and their influence on team processes as well as team performance has been informed by the current state of research on teams (Cannon-Bowers & Salas, 2001; Van den Bossche et al., 2011). However, the conceptual model includes a strong cognitive perspective and therefore does not emphasize affective, social, communication, and process-oriented components (Bartelt, Dennis, Yuan, & Barlow, 2013; Wildman et al., 2012). Hence, future work should advance the conceptual model and test the underlying theoretical assumption within controlled experimental settings.

Second, although our sample was large enough to achieve statistically significant results, the explained variance for our regression models was rather moderate. This indicates that besides the tested variables, others may have influenced the outcomes that were not tested in this study.

Third, the sample included a select group of participants from one university all enrolled in a specific course, thus prohibiting generalizations of results. Further, all participants of this study were inexperienced within the subject domains. This fact clearly limits the external validity of our findings (Campbell & Stanley, 1963). Accordingly, future studies should include various levels of difficulty, task type, and dispositions of participants within and across different subject domains.

Forth, although the participants' prior domain-specific knowledge was assessed, their causal, procedural, and metacognitive knowledge was not explicitly tested. Hence, possible effects of these variables were not addressed in this empirical investigation. Future studies may include a more comprehensive assessment of knowledge dimensions of individuals and teams (Wildman et al., 2012).

Fifth, the language-oriented analysis algorithm requires approximately 300 words for producing valid results (Pirnay-Dummer & Ifenthaler, 2011). Within this study, some teams produced texts under 300 words. A qualitative comparison of text with low word numbers and text with higher word numbers, however, did not reveal significant differences. Still, further studies are required to define the minimum word limit for valid AKOVIA analysis.

Sixth, the present study focused on the assessment of written text using German language. As AKOVIA currently supports German as well as English natural language processing, a wider

application using other languages is currently a clear limitation. Hence, a further development of AKOVIA should integrate several other language packages, for instance, those required in international large-scale studies. This would open up opportunities for further research focusing on the automated assessment of written texts in different languages. Such research would provide in-depth understanding of the validity of identical assessments in different languages and cultural contexts.

Seventh, individual team members were not assessed with regard to their subjective solution of the task. Accordingly, a future study may include an experimental variation where participants are asked to individually respond to the task. These individual solutions may then be aggregated for further analysis using the AKOVIA methodology (Ifenthaler & Pirnay-Dummer, 2014). A resulting research question may investigate the difference between aggregated solutions (based on individual responses) and team-created solutions.

Finally, a precise investigation of the learning-dependent progression of SMMs and their influence on team-based performance was not realized in this study. Also, information about the task and team shared knowledge was collected through self-report measures that have clear limitations with regard to their reliability and validity (Miyake, 1986). As research on the progression of SMMs is scarce (Barron, 2000), future studies may investigate the progression of IMMs and their influence on the progression of SMMs, and vice versa. Additionally, alternative assessment approaches for TaSMMs and TeSMMs may be investigated in order to better understand how these complex processes can be better assessed and optimized through interventions (Janssen, Erkens, Kirschner, & Kanselaar, 2010).

## Implications and Conclusion

The demand for computer-based assessment methodologies is evident through the current focus of international large-scale assessments such as PISA and PIAAC, as they enforce the implementation of alternative assessment environments. Clearly, there are numerous approaches for eliciting individual and team knowledge for various learning, teaching, and assessment purposes. For example, the recent introduction of mobile devices (e.g., tablets, PC) in K–12 education opens up new potentials for mobile learning environments, including augmented reality for learning, teaching, and assessment purposes (Cheng & Tsai, 2013; Ifenthaler & Eseryel, 2013). Another example is the rapid growth of online work environments in which teams gather, learn, and work together in virtual spaces, never meeting each other physically (Beer, Slack, & Armitt, 2005). Automated assessment and feedback systems for virtual environments may facilitate the shared understanding of tasks and project goals as well as provide insights into the team's progression toward common goals and successful solutions (Lenz & Machado, 2008).

However, most automated assessment approaches have not been tested for reliability and validity and are almost only applicable to single or small sets of data and specific subject domains (Fischer & Mandl, 2005; Greiff, Wüstenberg, Molnár, et al., 2013; Ruiz-Primo, Schultz, Li, & Shavelson, 2001). Therefore, new approaches are required that have not only been tested for reliability and validity but also provide a fast and economic way of analyzing larger sets of data (Greiff, Wüstenberg, Holt, Goldhammer, &



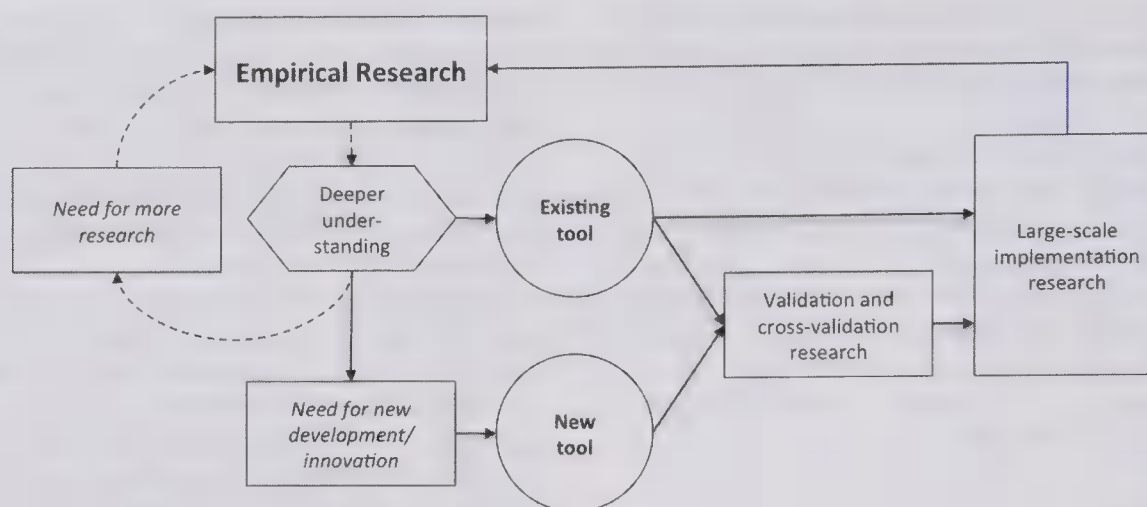


Figure 3. Research and methodology development framework.

Funke, 2013). Additionally, approaches for educational assessment also need to move beyond the perspective of correct and incorrect solutions (Mislevy et al., 2010). Hence, as we move further into the 21st century, the application of alternative assessment strategies is inevitable for current educational assessment (Savenye, 2014).

On the basis of the recent experience of developing, implementing, and empirically testing the AKOVIA methodology, a framework for integrating research and methodology development is outlined in Figure 3. This framework may both inform the innovation process (validation research) and improve the research without widening the risk for the research results. Until final acceptance of new methodologies such as AKOVIA, the standard tools are still used at that point. The triangulation will give interesting additional insight into research problems by means of post hoc analyses. Accordingly, the integration of methodological innovation alongside research standards and common research will shorten the time for implementation without harming the research process itself (Ifenthaler & Pirnay-Dummer, 2014). Without automation, many research projects would not be possible from the start. Hence, automation helps in raising the objectivity of outcomes and helps in realizing educational large-scale assessments. It also allows for a whole set of small and medium research projects to gain access to a reliable means of computer-based assessment (Csapó et al., 2012).

Given the recent developments in educational data mining and learning analytics (Long & Siemens, 2011), the automated assessment and analysis function of AKOVIA could be used to inform both decision makers (e.g., teachers, tutors, learning designers) and the learners during an ongoing learning process. Outcomes and results of these assessments could then be aggregated, transformed, and thus used to create feedback panels, dashboards, or even written feedback, based on the current learner and assessment model (Greller & Drachsler, 2012; Macfadyen & Dawson, 2012). Feedback on ongoing learning could be explicit by using results of the automated assessment and analysis (e.g., graphs, change indicators, as well as convergence toward a reference solution). Feedback could also be transformed for a more implicit use of the aggregations by implementing algorithms to create informative language-based feedback using quantitative measures of AKOVIA and linking them with a phrase database or badges.

Educational large-scale triangulation studies will continuously help to improve and converge the innovations in the methodology of assessment. To triangulate the use, usability, and feasibility of automated assessment methodologies, a large data set of a national and international magnitude may be preferable. This process may be repeated for validation and comparability should new methodologies be available or if existing methodologies change significantly. Still, challenges and critical issues in automated computer-based assessment, such as data security, accessibility, comparability, and compliance, need to be addressed before these systems are fully implemented in educational and work-related settings.

The scalability of the automated AKOVIA algorithm provides numerous applications for computer-based assessment environments within international large-scale assessments, such as collaborative problem solving, or teamwork. AKOVIA has the potential to widen the scope of current large-scale assessments and guide a new generation of cross-curricular natural language-oriented assessment environments.

## References

- Akkerman, S., Van den Bossche, P., Admiraal, W., Gijssels, W., Segers, M., Simons, R., & Kirschner, P. A. (2007). Reconsidering group cognition: From conceptual confusion to a boundary area between cognitive and socio-cultural perspectives? *Educational Research Review*, 2, 39–63. doi:10.1016/j.edurev.2007.02.001
- Al-Diban, S., & Ifenthaler, D. (2011). Comparison of two analysis approaches for measuring externalized mental models: Implications for diagnostics and applications. *Journal of Educational Technology & Society*, 14, 16–30.
- Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four process architecture. *Journal of Technology, Learning, and Assessment*, 1, 3–63.
- Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *I-S-T 2000 R*. Göttingen, Germany: Hogrefe.
- Badke-Schaub, P., Neumann, A., & Lauche, K. (2011). An observation-based method for measuring the sharedness of mental models in teams. In M. Boos, M. O. Kolbe, P. M. Kappeler, & T. Ellwart (Eds.), *Coordination in human and primate groups* (pp. 177–197). Berlin, Germany: Springer. doi:10.1007/978-3-642-15355-6\_10

- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84, 191–215. doi:10.1037/0033-295X.84.2.191
- Bandura, A. (1986). *Social foundation of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.
- Barron, B. (2000). Achieving coordination in collaborative problem-solving groups. *Journal of the Learning Sciences*, 9, 403–436. doi:10.1207/S15327809JLS0904\_2
- Bartelt, V. L., Dennis, A. R., Yuan, L., & Barlow, J. B. (2013). Individual priming in virtual team decision-making. *Group Decision and Negotiation*, 22, 873–896. doi:10.1007/s10726-012-9333-3
- Beer, M., Slack, F., & Armit, G. (2005). Collaboration and teamwork: Immersion and presence in an online learning environment. *Information Systems Frontiers*, 7, 27–37. doi:10.1007/s10796-005-5336-9
- Bierhals, R., Schuster, I., Kohler, P., & Badke-Schaub, P. (2007). Shared mental models—Linking team cognition and performance. *CoDesign*, 3, 75–94. doi:10.1080/15710880601170891
- Brill, E. (1995). *Unsupervised learning of disambiguation rules for part of speech tagging*. Paper presented at the Second Workshop on Very Large Corpora, WVLC-95, Boston, MA.
- Burke, C. S., Fiore, S. M., & Salas, E. (2003). The role of shared cognition in enabling shared leadership and team adaptability. In C. L. Pearce & J. A. Conger (Eds.), *Shared leadership: Reframing the hows and whys of leadership* (pp. 103–122). Thousand Oaks, CA: Sage. doi:10.4135/9781452229539.n5
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.
- Cannon-Bowers, J. A., & Salas, E. (2001). Reflections on shared cognition. *Journal of Organizational Behavior*, 22, 195–202. doi:10.1002/job.82
- Cannon-Bowers, J. A., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision making. In J. Castellan (Ed.), *Individual and group decision making: Current issues* (pp. 221–246). Mahwah, NJ: Erlbaum.
- Carley, K. M. (1997). Extracting team mental models through textual analysis. *Journal of Organizational Behavior*, 18, 533–558. doi:10.1002/(SICI)1099-1379(199711)18:1+<533::AID-JOB906>3.3.CO;2-V
- Chambers, B., & Abrami, P. C. (1991). The relationship between student team learning outcomes and achievement, causal attributions, and affect. *Journal of Educational Psychology*, 83, 140–146. doi:10.1037/0022-0663.83.1.140
- Cheng, K.-H., & Tsai, C.-C. (2013). Affordances of augmented reality in science learning: Suggestions for future research. *Journal of Science Education and Technology*, 22, 449–462. doi:10.1007/s10956-012-9405-9
- Chomsky, N. (1970). *Sprache und Geist*. Frankfurt am Main, Germany: Suhrkamp.
- Chung, G. K. W., O'Neil, H. F., & Herl, H. E. (1999). The use of computer-based collaborative knowledge mapping to measure team processes and team outcomes. *Computers in Human Behavior*, 15, 463–493. doi:10.1016/S0747-5632(99)00032-1
- Clariana, R. B. (2010). Deriving individual and group knowledge structure from network diagrams and from essays. In D. Ifenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 117–130). New York, NY: Springer. doi:10.1007/978-1-4419-5662-0\_7
- Cohen, P. R., Levesque, H. J., & Smith, I. A. (1997). On team formation. In G. Holmström-Hintikka & R. Tuomela (Eds.), *Contemporary action theory: Social action* (Vol. 2, pp. 87–114). Amsterdam, the Netherlands: Springer.
- Cooke, N. J., Salas, E., Cannon-Bowers, J. A., & Stout, R. J. (2000). Measuring team knowledge. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 42, 151–173. doi:10.1518/001872000779656561
- Cooke, N. J., Salas, E., Kiekel, P. A., & Bell, B. (2004). Advances in measuring team cognition. In E. Salas & S. M. Fiore (Eds.), *Team cognition: Understanding the factors that drive process and performance* (pp. 83–106). Washington, DC: American Psychological Association. doi:10.1037/10690-005
- Coronges, K. A., Stacy, A. W., & Valente, T. W. (2007). Structural comparison of cognitive associative networks in two populations. *Journal of Applied Social Psychology*, 37, 2097–2129. doi:10.1111/j.1559-1816.2007.00253.x
- Cox, T., & Blake, S. (1991). Managing cultural diversity: Implications for organizational competitiveness. *Academy of Management Executive*, 5, 45–56.
- Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 143–230). Amsterdam, the Netherlands: Springer. doi:10.1007/978-94-007-2324-5\_4
- Ding, Y. (2001). A review of ontologies with the semantic web in view. *Journal of Information Science*, 27, 377–384. doi:10.1177/016555150102700603
- Eccles, D. W., & Tenenbaum, G. (2004). Why an expert team is more than a team of experts: A social-cognitive conceptualization of team coordination and communication in sport. *Journal of Sport & Exercise Psychology*, 26, 542–560.
- Ellson, J., Gansner, E. R., Koutsofios, E., North, S. C., & Woodhull, G. (2003). *GraphViz and Dynagraph. Static and dynamic graph drawing tools*. Florham Park, NJ: AT&T Labs.
- Fischer, F., Bruhn, J., Gräsel, C., & Mandl, H. (2002). Fostering collaborative knowledge construction with visualization tools. *Learning and Instruction*, 12, 213–232. doi:10.1016/S0959-4752(01)00005-6
- Fischer, F., & Mandl, H. (2005). Knowledge convergence in computer-supported collaborative learning: The role of external representation tools. *Journal of the Learning Sciences*, 14, 405–441. doi:10.1207/s15327809jls1403\_3
- Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25, 285–307. doi:10.1080/01638539809545029
- Frazier, L. (1999). *On sentence interpretation*. Dordrecht, the Netherlands: Kluwer. doi:10.1007/978-94-011-4599-2
- Frey, A., & Hartig, J. (2013). Wann sollten computerbasierte Verfahren zur Messung von Kompetenzen anstelle von papier- und bleistift-basierten Verfahren eingesetzt werden? [In which settings should computer-based tests be used instead of paper and pencil-based tests?] *Zeitschrift für Erziehungswissenschaft*, 16, 53–57. doi:10.1007/s11618-013-0385-1
- Gentner, D., & Stevens, A. L. (1983). *Mental models*. Hillsdale, NJ: Erlbaum.
- Goldsmith, T. E., & Davenport, D. M. (1990). Assessing structural similarity of graphs. In R. W. Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization* (pp. 75–87). Norwood, NJ: Ablex.
- Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F., & Funke, J. (2013). Computer-based assessment of complex problem solving: Concept, implementation, and application. *Educational Technology Research and Development*, 61, 407–421. doi:10.1007/s11423-013-9301-x
- Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts - Something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology*, 105, 364–379. doi:10.1037/a0031856
- Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Educational Technology & Society*, 15, 42–57.
- Guzzo, R. A., & Dickson, M. W. (1996). Teams in organizations: Recent research on performance and effectiveness. *Annual Review of Psychology*, 47, 307–338. doi:10.1146/annurev.psych.47.1.307



- Gweon, G., Rosé, C. P., Wittwer, J., & Nueckles, M. (2005). Supporting efficient and reliable content analysis using automatic text processing technology. In M. F. Costabile & F. Paternò (Eds.), *Human-computer interaction - INTERACT 2005* (Vol. 3585, pp. 1112–1115). Berlin, Germany: Springer. doi:10.1007/11555261\_117
- Hambrick, D. C., Cho, T. S., & Chen, M. J. (1996). The influence of top management team heterogeneity on firms' competitive moves. *Administrative Science Quarterly*, 41, 659–684. doi:10.2307/2393871
- Helbig, H. (2006). *Knowledge representation and the semantics of natural language*. Berlin, Germany: Springer.
- Herl, H. E., O'Neil, H. F., Chung, G. K. W. K., & Schacter, J. (1999). Reliability and validity of a computer-based knowledge mapping system to measure content understanding. *Computers in Human Behavior*, 15, 315–333. doi:10.1016/S0747-5632(99)00026-6
- Horwitz, S. K. (2005). The compositional impact of team diversity on performance: Theoretical considerations. *Human Resource Development Review*, 4, 219–245. doi:10.1177/1534484305275847
- Horwitz, S. K., & Horwitz, I. B. (2007). The effects of team diversity on team outcomes: A meta-analytic review of team demography. *Journal of Management*, 33, 987–1015. doi:10.1177/0149206307308587
- Ifenthaler, D. (2009). Model-based feedback for improving expertise and expert performance. *Technology, Instruction, Cognition and Learning*, 7, 83–101.
- Ifenthaler, D. (2010a). Bridging the gap between expert-novice differences: The model-based feedback approach. *Journal of Research on Technology in Education*, 43, 103–117.
- Ifenthaler, D. (2010b). Relational, structural, and semantic analysis of graphical representations and concept maps. *Educational Technology Research and Development*, 58, 81–97. doi:10.1007/s11423-008-9087-4
- Ifenthaler, D. (2010c). Scope of graphical indices in educational diagnostics. In D. Ifenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 213–234). New York, NY: Springer. doi:10.1007/978-1-4419-5662-0\_12
- Ifenthaler, D. (2011). Identifying cross-domain distinguishing features of cognitive structures. *Educational Technology Research and Development*, 59, 817–840. doi:10.1007/s11423-011-9207-4
- Ifenthaler, D. (2012). Determining the effectiveness of prompts for self-regulated learning in problem-solving scenarios. *Journal of Educational Technology & Society*, 15, 38–52.
- Ifenthaler, D., & Eseryel, D. (2013). Facilitating complex learning by mobile augmented reality learning environments. In R. Huang, Kinshuk, & J. M. Spector (Eds.), *Reshaping learning: The frontiers of learning technologies in a global context* (pp. 415–438). New York, NY: Springer.
- Ifenthaler, D., & Lehmann, T. (2012). Preactional self-regulation as a tool for successful problem solving and learning. *Technology, Instruction, Cognition and Learning*, 9, 97–110.
- Ifenthaler, D., Masduki, I., & Seel, N. M. (2011). The mystery of cognitive structure and how we can detect it: Tracking the development of cognitive structures over time. *Instructional Science*, 39, 41–61. doi:10.1007/s11251-009-9097-6
- Ifenthaler, D., & Pirnay-Dummer, P. (2014). Model-based tools for knowledge assessment. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (4th ed., pp. 289–301). New York, NY: Springer.
- Ifenthaler, D., & Seel, N. M. (2013). Model-based reasoning. *Computers and Education*, 64, 131–142. doi:10.1016/j.compedu.2012.11.014
- Janssen, J., Erkens, G., Kirschner, P. A., & Kanselaar, G. (2010). Effects of representational guidance during computer-supported collaborative learning. *Learning and Instruction*, 38, 59–88. doi:10.1007/s11251-008-9078-1
- Johnson, T. E., Ifenthaler, D., Pirnay-Dummer, P., & Spector, J. M. (2009). Using concept maps to assess individuals and teams in collaborative learning environments. In P. L. Torres & R. C. V. Marriott (Eds.), *Handbook of research on collaborative learning using concept mapping* (pp. 358–381). Hershey, PA: Information Science Publishing. doi:10.4018/978-1-59904-992-2.ch018
- Johnson, T. E., & Lee, Y. (2008). The relationship between shared mental models and task performance in an online team-based learning environment. *Performance Improvement Quarterly*, 21, 97–112. doi:10.1002/piq.20033
- Johnson, T. E., Lee, Y., Lee, M., & O'Connor, D. L. (2007). Measuring sharedness of team-related knowledge: Design and validation of a shared mental model instrument. *Human Resource Development International*, 10, 437–454. doi:10.1080/13678860701723802
- Johnson, T. E., Pirnay-Dummer, P., Ifenthaler, D., Mendenhall, A., Karaman, S., & Tennenbaum, G. (2011). Text summaries or concept maps: Which better represent reading text conceptualization? *Technology, Instruction, Cognition and Learning*, 8, 297–312.
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, United Kingdom: Cambridge University Press.
- Johnson-Laird, P. N. (1989). Mental models. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 469–499). Cambridge, MA: MIT Press.
- Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge*. Hillsdale, NJ: Erlbaum.
- Jonassen, D. H., & Cho, Y. H. (2008). Externalizing mental models with mindtools. In D. Ifenthaler, P. Pirnay-Dummer, & J. M. Spector (Eds.), *Understanding models for learning and instruction: Essays in honor of Norbert M. Seel* (pp. 145–159). New York, NY: Springer. doi:10.1007/978-0-387-76898-4\_7
- Kanaga, K., & Kossler, M. E. (2011). *How to form a team: Five keys to high performance*. Greensboro, NC: Center for Creative Leadership.
- Kant, I. (1998). *Kritik der reinen Vernunft* [Critique of pure reason]. Hamburg, Germany: Meiner Verlag. (Original work published 1781)
- Katzenbach, J. R., & Smith, D. K. (1993). The discipline of teams. *Harvard Business Review*, 71, 111–120.
- Katzenbach, J. R., & Smith, D. K. (2003). *The wisdom of teams: Creating the high-performance organization*. New York, NY: Collins Business.
- Kim, M. K. (2012). Cross-validation study of methods and technologies to assess mental models in a complex problem solving situation. *Computers in Human Behavior*, 28, 703–717. doi:10.1016/j.chb.2011.11.018
- Klimoski, R., & Mohammed, S. (1994). Team mental model: Construct or metaphor? *Journal of Management*, 20, 403–437. doi:10.1016/0149-2063(94)90021-3
- Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest*, 7, 77–124.
- Lenz, R., & Machado, C. (2008). Virtual teamwork: A product of globalization. In N. P. Barsky, M. Clements, J. Ravn, & K. Smith (Eds.), *The power of technology for learning* (pp. 77–93). Amsterdam, the Netherlands: Springer. doi:10.1007/978-1-4020-8747-9\_5
- Lim, B.-C., & Klein, K. J. (2006). Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy. *Journal of Organizational Behavior*, 27, 403–418. doi:10.1002/job.387
- Long, P. D., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, 46, 31–40.
- Macfadyen, L., & Dawson, S. (2012). Numbers are not enough. Why e-Learning analytics failed to inform an institutional strategic plan. *Educational Technology & Society*, 15, 149–163.
- Mandl, H., & Fischer, F. (2000). Mapping-Techniken und Begriffsnetze in Lern- und Kooperationsprozessen [Mapping techniques and concept maps in learning and cooperating processes]. In H. Mandl & F. Fischer (Eds.), *Wissen sichtbar machen – Wissensmanagement mit Mapping-Techniken* [Making knowledge visible: Knowledge management with mapping techniques] (pp. 3–12). Göttingen, Germany: Hogrefe.

- Markman, A. B. (1999). *Knowledge representation*. Mahwah, NJ: Erlbaum.
- Marks, M. A., Zaccaro, S. J., & Mathieu, J. E. (2000). Performance implications of leader briefings and team-interaction training for team adaptation to novel environments. *Journal of Applied Psychology*, 85, 971–986. doi:10.1037/0021-9010.85.6.971
- Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Cannon-Bowers, J. A., & Salas, E. (2005). Scaling the quality of teammates' mental models: Equifinality and normative comparisons. *Journal of Organizational Behavior*, 26, 37–56. doi:10.1002/job.296
- Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology*, 85, 273–283. doi:10.1037/0021-9010.85.2.273
- McKeown, J. O. (2009). *Using annotated concept map assessments as predictors of performance and understanding of complex problems for teacher technology integration*. Tallahassee: Florida State University.
- Mesmer-Magnus, J. R., & Dechurch, L. A. (2009). Information sharing and team performance: A meta-analysis. *Journal of Applied Psychology*, 94, 535–546. doi:10.1037/a0013773
- Minsky, M. (1981). A framework for representing knowledge in mind design. In R. J. Brachmann & H. J. Levesque (Eds.), *Readings in knowledge representation* (pp. 245–262). Los Altos, CA: Morgan Kaufmann.
- Mislevy, R. J., Behrens, J. T., Bennett, R. E., Demark, S. F., Frezzo, D. C., Levy, R., . . . Winters, F. I. (2010). On the roles of external knowledge representations in assessment design. *Journal of Technology, Learning, and Assessment*, 8, 4–55.
- Miyake, N. (1986). Constructive interaction and the iterative process of understanding. *Cognitive Science*, 10, 151–177. doi:10.1207/s15516709cog1002\_2
- Mohammed, S., & Dumville, B. C. (2001). Team mental models in a team knowledge framework: Expanding theory and measurement across disciplinary boundaries. *Journal of Organizational Behavior*, 22, 89–106. doi:10.1002/job.86
- Mohammed, S., Klimoski, R., & Rentsch, J. R. (2000). The measurement of team mental models: We have no shared schema. *Organizational Research Methods*, 3, 123–165. doi:10.1177/109442810032001
- Moreland, R. L., & Levine, J. M. (1992). The composition of small groups. In E. Lawler, B. Markovsky, C. Ridgeway, & H. Walker (Eds.), *Advances in group processes* (pp. 237–280). Greenwich, CT: JAI Press.
- O'Neil, H. F., Chuang, S.-H., & Baker, E. L. (2010). Computer-based feedback for computer-based collaborative problem solving. In D. Ifenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 261–280). New York, NY: Springer. doi:10.1007/978-1-4419-5662-0\_14
- Organisation for Economic Co-Operation and Development. (2010). *PISA computer-based assessment of student skills in science*. Paris, France: Author.
- Organisation for Economic Co-Operation and Development. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris, France: Author.
- Piray-Dummer, P., & Ifenthaler, D. (2010). Automated knowledge visualization and assessment. In D. Ifenthaler, P. Piray-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 77–115). New York, NY: Springer. doi:10.1007/978-1-4419-5662-0\_6
- Piray-Dummer, P., & Ifenthaler, D. (2011). Reading guided by automated graphical representations: How model-based text visualizations facilitate learning in reading comprehension tasks. *Instructional Science*, 39, 901–919. doi:10.1007/s11251-010-9153-2
- Piray-Dummer, P., Ifenthaler, D., & Spector, J. M. (2010). Highly integrated model assessment technology and tools. *Educational Technology Research and Development*, 58, 3–18. doi:10.1007/s11423-009-9119-8
- Pollio, H. R. (1966). *The structural basis of word association behavior*. The Hague, the Netherlands: Mouton.
- Ruiz-Primo, M. A., Schultz, S. E., Li, M., & Shavelson, R. J. (2001). Comparison of the reliability and validity of scores from two concept-mapping techniques. *Journal of Research in Science Teaching*, 38, 260–278. doi:10.1002/1098-2736(200102)38:2<260::AID-TEA1005>3.0.CO;2-F
- Rumelhart, D. E., & Norman, D. A. (1978). Accretion, tuning and restructuring: Three models of learning. In R. L. Klatzky & J. W. Cotton (Eds.), *Semantic factors in cognition* (pp. 37–53). Hillsdale, NJ: Erlbaum.
- Salas, E., Cooke, N. J., & Rosen, M. A. (2008). On teams, teamwork, and team performance: Discoveries and developments. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50, 540–547. doi:10.1518/001872008X288457
- Salas, E., Dickinson, T. L., Converse, S. A., & Tannenbaum, S. I. (1992). Toward an understanding of team performance and training. In R. W. Swezey & E. Salas (Eds.), *Teams: Their training and performance* (pp. 3–29). Westport, CT: Ablex.
- Savenye, W. C. (2014). Perspectives on assessment of educational technologies for informal learning. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 257–267). New York, NY: Springer. doi:10.1007/978-1-4614-3185-5\_21
- Schvaneveldt, R. W. (Ed.). (1990). *Pathfinder associative networks: Studies in knowledge organization*. Norwood, NJ: Ablex.
- Searle, J., & Grewendorf, G. (2002). *Speech acts, mind, and social reality*. Dordrecht, the Netherlands: Kluwer Academic Publishers.
- Seel, N. M. (1999). Educational diagnosis of mental models: Assessment problems and technology-based solutions. *Journal of Structural Learning and Intelligent Systems*, 14, 153–185.
- Sikorski, E. G., Johnson, T. E., & Ruscher, P. H. (2012). Team knowledge sharing intervention effects on team shared mental models and student performance in an undergraduate science course. *Journal of Science Education and Technology*, 21, 641–651. doi:10.1007/s10956-011-9353-9
- Spector, J. M., & Koszalka, T. A. (2004). *The DEEP methodology for assessing learning in complex domains* (Final report to the National Science Foundation Evaluative Research and Evaluation Capacity Building). Syracuse, NY: Syracuse University.
- Strasser, A. (2010). A functional view toward mental representations. In D. Ifenthaler, P. Piray-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 15–25). New York, NY: Springer. doi:10.1007/978-1-4419-5662-0\_2
- Taricani, E. M., & Clariana, R. B. (2006). A technique for automatically scoring open-ended concept maps. *Educational Technology Research and Development*, 54, 65–82. doi:10.1007/s11423-006-6497-z
- Tittmann, P. (2010). Graphs and networks. In D. Ifenthaler, P. Piray-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 177–188). New York, NY: Springer. doi:10.1007/978-1-4419-5662-0\_10
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327–352. doi:10.1037/0033-295X.84.4.327
- Van den Bossche, P., Gijssels, W., Segers, M., Woltjer, G., & Kirschner, P. A. (2011). Team learning: Building shared mental models. *Instructional Science*, 39, 283–301. doi:10.1007/s11251-010-9128-3
- Wildman, J. L., Thayer, A. L., Pavlas, D., Salas, E., Stewart, J. E., & Howse, a. W. R. (2012). Team knowledge research: Emerging trends and critical needs. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 54, 84–111. doi:10.1177/0018720811425365
- Wittgenstein, L. (1922). *Tractatus logico-philosophicus*. New York, NY: Harcourt Brace & Company.

Received April 15, 2013

Revision received October 11, 2013

Accepted October 17, 2013 ■



# The Computer-Based Assessment of Complex Problem Solving and How It Is Influenced by Students' Information and Communication Technology Literacy

Samuel Greiff, André Kretzschmar,  
and Jonas C. Müller  
University of Luxembourg

Birgit Spinath  
University of Heidelberg

Romain Martin  
University of Luxembourg

The 21st-century work environment places strong emphasis on nonroutine transversal skills. In an educational context, complex problem solving (CPS) is generally considered an important transversal skill that includes knowledge acquisition and its application in new and interactive situations. The dynamic and interactive nature of CPS requires a computer-based administration of CPS tests such that the assessment of CPS might be partially confounded with information and communication technology (ICT) literacy. To establish CPS as a distinct construct that involves complex cognitive processes not covered by other general cognitive abilities and not related to ICT literacy, it is necessary to investigate the influence of ICT literacy on CPS and on the power of CPS to predict external educational criteria. We did so in 3 different samples of either high school or university students using a variety of instruments to measure ICT literacy and general cognitive ability. Convergent results based on structural equation modeling and confirmatory factor analyses across the studies showed that ICT literacy was weakly or moderately related to CPS, and these associations were similar to those between ICT and other general cognitive abilities. Furthermore, the power of CPS to predict external educational criteria over and above general cognitive ability remained even if the influence of ICT literacy on CPS was controlled for. We conclude that CPS is a distinct construct that captures complex cognitive processes not generally found in other assessments of general cognitive ability or of ICT literacy.

**Keywords:** ICT literacy, complex problem solving, reasoning, working memory, MicroDYN

**Supplemental materials:** <http://dx.doi.org/10.1037/a0035426.supp>

Occupational demands are changing rapidly in the 21st century (Autor & Dorn, 2009; Organisation for Economic Co-Operation and Development [OECD], 2010; Spitz-Oener, 2006), and nonroutine skills are becoming more and more

important, whereas the importance of routine skills that are characterized by the repetitive occurrence of similar situations is decreasing (Autor, Levy, & Murnane, 2003). As nonroutine skills can be used in many situations and for different problems, these skills are by definition transversal rather than domain specific. Furthermore, facilitating transversal competencies is a central objective in a number of educational programs (Mayer & Wittrock, 2006), and transversal skills such as complex problem solving (CPS) play an important role in everyday life (Funke, 2010). Although transversal skills are found in a number of areas and encompass skills such as metacognition, creativity, as well as collaborative and CPS, this last skill is a particularly important and promising transversal skill that has recently received a lot of attention, especially in educational contexts. For instance, Mayer and Wittrock (2006) stated that one of education's greatest challenges is making students good problem solvers. Therefore, it is not surprising that CPS is now an integral part of international educational large-scale assessments such as the Programme for International Student Assessment (PISA), arguably the most influential educational large-scale survey worldwide (OECD, 2009a). According to Buchner (Frensch & Funke, 1995), CPS can be defined as:

---

This article was published Online First February 17, 2014.

Samuel Greiff, André Kretzschmar, and Jonas C. Müller, Educational Measurement and Applied Cognitive Science (EMACS) research unit, University of Luxembourg, Luxembourg; Birgit Spinath, Department of Psychology, University of Heidelberg, Heidelberg, Germany; Romain Martin, Educational Measurement and Applied Cognitive Science (EMACS) research unit, University of Luxembourg, Luxembourg.

This research was funded by Luxembourgish Research Foundation Grant FNR ATTRACT ASSKI21, European Union Grant 290683 (LLLight in Europe), German Research Foundation Grant DFG Fu 173-14/1, and German Federal Ministry of Education and Research Grant LSA004. We are grateful to the Technology Based Assessment group at DIPF (<http://tba.dipf.de>) for providing the authoring tool CBA Item Builder and technical support.

Correspondence concerning this article should be addressed to Samuel Greiff, EMACS research unit, University of Luxembourg, 6, rue Richard Coudenhove-Kalergi, 1359 Luxembourg-Kirchberg, Luxembourg. E-mail: [samuel.greiff@uni.lu](mailto:samuel.greiff@uni.lu)

The successful interaction with task environments that are dynamic (i.e., change as a function of user's intervention and/or as a function of time) and in which some, if not all, of the environment's regularities can only be revealed by successful exploration and integration of the information gained in that process. (p. 14)

CPS has some unique characteristics that distinguish it from other abilities such as reasoning or working memory (cf. Dörner, Kreuzig, Reither, & Stäudel, 1983; Fischer, Greiff, & Funke, 2012; Funke, 2001). It refers to complex and nontransparent situations because not all of the necessary information to solve the problem is available until the problem solver interacts with the problem dynamically. That is, some information is hidden at the outset (Frensch & Funke, 1995). Furthermore, dynamic changes and highly interrelated elements in CPS require problem solvers to actively generate information by applying adequate strategies. Finally, multiple goals have to be taken into account when trying to solve a problem. Thus, a dynamic interaction between the problem solver and the task situation is an inherent feature of CPS (Wirth & Klieme, 2003), and this kind of interaction is not conceptually inherent to other cognitive abilities. These characteristics of CPS are typical of transversal skills, and this is why CPS has a prominent relevance among them.

In general, CPS is composed of two overarching processes: the active acquisition of knowledge about a problem situation (knowledge acquisition; Mayer & Wittrock, 2006) and the active use of this knowledge, that is, finding a solution to a problem (knowledge application; Novick & Bassok, 2005). According to the definition of CPS and the aforementioned characteristics, especially those of interactivity and dynamics, the assessment of CPS should be particularly fruitful in the context of computer-based assessment (CBA). CBA provides a unique assessment environment for the required dynamic and interactive situations that cannot be provided by the use of paper-and-pencil instruments (Kyllonen, 2009; Williamson, Mislevy, & Bejar, 2006).

In the area of assessment, there has always been high interest in CPS as a higher order thinking skill (Kuhn, 2009) that may both conceptually and empirically complement assessments of other general cognitive abilities such as reasoning, working memory, perceptual speed, and so forth. In fact, recent findings have suggested that CPS has an added value beyond other general cognitive abilities. For example, an added value of CPS above and beyond reasoning abilities has been found in predictions of academic achievement (e.g., Greiff et al., 2013; Greiff, Wüstenberg, & Funke, 2012; Wüstenberg, Greiff, & Funke, 2012) and supervisor ratings of professional success (Danner, Hagemann, Schankin, Hager, & Funke, 2011). In general, it is assumed that the underlying cognitive processes of CPS are correlated with and yet distinct from other general cognitive abilities (cf. Schweizer, Wüstenberg, & Greiff, 2013; Wüstenberg et al., 2012) and that in addition, more complex cognitive processes related to knowledge acquisition and knowledge application are responsible for the added value of CPS in predicting relevant external criteria (cf. Gonzalez, Thomas, & Vanyukov, 2005; Greiff, Wüstenberg, et al., 2013; Wenke, Frensch, & Funke, 2005). However, a final embedding of CPS in the nomological network of theories on cognitive ability (e.g., in the Cattell-Horn-Carroll [CHC] theory; McGrew, 2009) has not yet been fully accomplished (cf. Greiff, Wüstenberg, et al., 2013).

## ICT Literacy and How It Might Be Related to CPS

In recent studies that have demonstrated the incremental validity of CPS beyond other cognitive abilities, it has been argued that there are unique characteristics and complex cognitive processes inherent in CPS and that these are not found in the conceptualizations of other general cognitive abilities (Raven, 2000). However, due to the fact that CPS assessment instruments require computer-based test administration, researchers cannot rule out the possibility that the added value of CPS may stem from an influence of computer literacy on CPS test results. In this line of thinking, CPS tests would then provide an indirect measure of information and communication technology (ICT) literacy.

Due to the enhanced complexity and attractiveness of CBAs, it has been assumed that ICT literacy might have a strong impact on performance in CBAs, especially if these assessments require more complex interactions with the computer, a requirement that holds in particular for CPS. In a comprehensive definition, Tsai (2002) described ICT literacy as "the basic knowledge, skills, and attitudes needed by all citizens to be able to deal with computer technology in their daily life" (p. 69). Thus, declarative, procedural, and attitudinal aspects are covered by this conceptualization, which indicates that it is not only computer knowledge and skills that are important for handling CBAs. Affective components can influence performance as well. For instance, high computer anxiety may lead to discomfort when using the computer, resulting in lower performance on exploratory behavior. Thus, the added value of CPS above and beyond general cognitive ability could be due to tests of CPS inadvertently providing an indirect assessment of ICT literacy rather than to CPS representing additional complex cognitive processes required by the problem-solving situation. As a consequence, the overall validity of the CPS construct as a complex cognitive skill that is distinct from other general cognitive abilities might be threatened because empirical access to this construct is inevitably bound to the computer-based administration mode (cf. Parshall, Spray, Kalohn, & Davey, 2002; Russell, Goldberg, & O'Connor, 2003).

This concern is an important one especially against the background of the general shift from paper-pencil tests toward CBA (Goldhammer, Naumann, & Keßel, 2013). For example, early large-scale assessments (e.g., the PISA survey in 2003) used only paper-and-pencil assessments. Computer-based testing was partly introduced in PISA 2006 (OECD, 2007), substantially extended in PISA 2012 (OECD, 2010), included in the Programme for the International Assessment of Adult Competencies (OECD, 2009b), and, finally, will constitute the major mode of delivery in PISA 2015 (OECD, 2012). This progression can be accounted for by several general advantages of CBA (cf. Scheuermann & Björns-son, 2009; Van der Linden & Glas, 2000) such as high standardization and test efficiency, the logging of behavioral and process data, the possibility of automatic scoring, and the application of adaptive testing.

As any computer-administered test requires the test taker to interact with the computer, the influence of ICT literacy can be considered a general threat to the validity of any construct assessed via this mode of administration; thus, the issue is not limited to CPS. Studies concerning mode-of-delivery effects in general have addressed this topic but have provided inconsistent results. In an older meta-analysis, Mead and Drasgow (1993) found no overall



difference between paper-pencil and CBAs. Nevertheless, the authors warned against drawing the conclusion that there is no test mode effect and thus no influence of ICT literacy at all. Therefore, it is not surprising that several studies have reported relevant test-mode effects (for an overview, see Clariana & Wallace, 2002; Russell et al., 2003). These inconsistent results might also be due to the fact that different types of tests imply more or less complex interactions with the computer and thus require different levels of ICT literacy.

In this sense, constructs such as CPS, which are based on more innovative item types that reflect the dynamic interaction and display features that are offered by the computer, are also more prone to the undesirable influence of ICT literacy. At the same time, for tests that rely on these new item types such as CPS, it is not possible to conduct studies on test-mode administration effects because the classical paper-and-pencil mode of administration is simply not available for these item types and thus cannot be compared with the computer-administrated mode. For other general cognitive abilities such as reasoning and working memory, empirical studies can be conducted to examine how assessment may be affected by changing the mode of delivery from paper-and-pencil to computer based. The question then arises: How can researchers understand and quantify the influence of ICT literacy on an assessment of complex cognitive skills such as CPS when the assessment can be administered only on the computer (Kyllonen, 2009)? That is, for CPS, it is yet unclear whether its added value in predicting external criteria (e.g., Schweizer et al., 2013; Wüstenberg et al., 2012) originates from the indirect assessment of ICT literacy or from the assessment of additional and relevant cognitive processes as mentioned above and as conceptually assumed.

### **The Added Value of CPS: Cognitive Processes or Merely ICT Literacy?**

The central question of the current study is about the influence of ICT literacy on the CBA of CPS. Specifically, we asked how we could explain the added value in terms of the incremental validity of CPS in predicting relevant external criteria above and beyond general cognitive ability. We proposed two conspicuous explanations: additional cognitive processes, on the one hand, and additional demands on students' ICT literacy, on the other hand. Establishing the construct of CPS as a transversal skill would be warranted and an assessment of CPS in international large-scale studies would be justified only if the first explanation were to hold.

To tackle this question, we had to examine the simultaneous influence of general cognitive ability and ICT literacy on CPS. In general, cognitive abilities and CPS share cognitive processes to a certain extent (cf. Greiff, Wüstenberg, et al., 2013; Wüstenberg et al., 2012). However, according to the definition and characteristics of CPS, unique processes are supposed to be inherent to CPS. Different from general cognitive ability, which mainly requires a mere sequence of simple cognitive processes, CPS requires a series of different cognitive processes such as action planning and implementing, strategic development, knowledge acquisition, and self-regulation (Funke, 2010; Raven, 2000).

As outlined above, ICT literacy might also have an influence on CPS as a consequence of the mode of delivery. Each CBA requires at least basic computer knowledge as well as the related perceptual

and motor skills that are needed to use the computer interface. Furthermore, by definition, higher ICT literacy leads to a more familiar and intuitive handling of the computer interface, or to look at it the other way around, if a student's ICT literacy in an assessment context is very low, cognitive resources have to be used to understand the computer interface, for example. This would tie up a large amount of cognitive capacity that would then not be available for CPS even if the core interest of the assessment lies in CPS (Goldhammer et al., 2013; see also cognitive load theory; Sweller, 2005). In conclusion, to determine whether CPS is more than general cognitive ability and ICT literacy combined, the relations of both of them to CPS must be examined simultaneously.

Surprisingly, there are hardly any empirical findings with regard to the impact of ICT literacy on CPS and none at all with regard to the relations between ICT literacy, general cognitive ability, and CPS. Hartig and Klieme (2005) reported small relations between CPS and self-reported ICT literacy. Further, Süß (1996) reported moderate to high correlations between objective indicators of ICT literacy and CPS. However, these early studies did not account for the development of innovative, more user-friendly computer interfaces or the substantial changes in the use and importance of computers in everyday life; these changes have thus resulted in study participants who can be considered "digital natives" (Prensky, 2001). In a recent study, Sonnleitner, Keller, Martin, and Brunner (2013) highlighted that an added value of CPS beyond reasoning is found only in academic achievement criteria that are assessed via computer but not in paper-pencil-assessed criteria. They concluded that the added value of CPS is merely an effect of test mode and thus of ICT literacy.

Overall, there are two possible explanations for the added value of CPS recently reported in the literature: complex cognitive processes that are not included in concepts of general cognitive ability or an indirect but substantial influence of ICT literacy on the CBA of CPS. However, there are very few studies that have targeted this issue, and these studies have produced inconsistent findings. The purpose of the current study was to take a deeper look into the impact of ICT literacy on CPS and on CBAs of transversal skills in general.

### **Purpose of the Study**

Generally speaking, we want to advance knowledge on the question of how CPS and ICT literacy are related to each other and whether CPS indeed yields a valuable marker of additional complex cognitive processes or whether it is a confounded indicator of general cognitive ability and ICT literacy. Thus, we addressed the question of whether the added value of CPS reported in some studies could be explained by ICT literacy or, in other words, whether CPS is something other than general cognitive abilities such as reasoning or working memory and ICT literacy combined. To this end, we derived three research questions.

Research Question 1: How are ICT literacy and CPS related to each other?

For the first question, we examined latent correlations between ICT literacy and the CPS dimensions of knowledge acquisition and knowledge application.

Research Question 2: Does ICT literacy more strongly predict a CBA of CPS than ICT literacy predicts the assessment of general cognitive ability?

In the next step, we analyzed the latent regression of CPS and general cognitive ability on ICT literacy and tested whether ICT literacy would predict CPS more strongly than it predicted general cognitive ability.

Research Question 3: Can the added value of CPS be explained by ICT literacy or are distinct cognitive processes in CPS responsible for the added value?

For the last question, we examined whether CPS could explain academic achievement above and beyond general cognitive ability and ICT literacy combined or whether controlling for general cognitive ability and ICT literacy would result in the nonsignificant prediction of external criteria such as academic achievement by CPS.

To answer these questions, we used three different and diverse samples containing both high school and university students. In all these samples, the added value of CPS beyond general cognitive ability has been shown to exist in previously published articles (Study A: Wüstenberg et al., 2012; Study B: Greiff, Fischer, et al., 2013; Study C: Schweizer et al., 2013). However, the added value of CPS beyond ICT literacy and general cognitive ability was not tested in any of these studies. Thus, we extended the original analyses by adding ICT literacy, which was operationalized in diverse ways. According to its definition (see above; Tsai, 2002), ICT literacy is a broad concept composed of cognitive and affective aspects. Thus, a valid assessment of ICT literacy needs to endorse different operationalizations and methods (Ballantine, McCourt Larres, & Oyeler, 2007; Goldhammer et al., 2013; Van Braak, 2004). Therefore, we used subjective self-reports and different objective performance tests. Consequently, we used different operationalizations of general cognitive ability as well: figural reasoning and working memory capacity. Finally, an acknowledged and well-validated measure of CPS, namely MicroDYN (Greiff et al., 2012), was used in all three samples. To sum up, this approach allowed us to examine our research questions in different samples with heterogeneous assessments of ICT literacy and general cognitive ability. Our findings can thus be cross-checked to ensure that they are replicable and generalizable (Brennan, 1983).

## Method

### Assessment Instrument for CPS: MicroDYN

In all three studies (Study A, Study B, and Study C), a set of tasks used in the MicroDYN approach (Greiff et al., 2012) was used to assess CPS. In MicroDYN, students are first asked to detect causal relations in a dynamic system composed of several input and output variables. Subsequently, they are asked to control the system. These two tasks directly relate to the two characteristic CPS dimensions introduced above, knowledge acquisition and knowledge application, thus ensuring the theoretical embedding of the MicroDYN approach.

Recent results have indicated that MicroDYN is a reliable (consistent Cronbach's  $\alpha$ s  $> .70$ ; cf. Greiff et al., 2012; Wüstenberg et al., 2012) and valid assessment instrument (Greiff, Wüstenberg, et al., 2013; Molnár, Greiff, & Csapó, 2013; Schweizer et al., 2013; Wüstenberg et al., 2012) that sufficiently reflects the theo-

retical concept of CPS. For instance, MicroDYN as an operationalization of CPS shows incremental validity in predicting academic achievement beyond general cognitive abilities such as reasoning and working memory (Greiff, Wüstenberg, et al., 2013; Schweizer et al., 2013; Wüstenberg et al., 2012). Further, a substantial number of the items used to assess CPS in 15-year-old students across a number of countries in the PISA 2012 study were developed within the MicroDYN approach.

A set of MicroDYN tasks typically encompasses five to 10 independent complex problems (also referred to as *microworlds* in the literature; cf. Funke, 2001), with time on task being approximately 5 min for each CPS task. Each task has an underlying causal structure unknown to the student and is divided into two subsequent phases: Phase 1, in which knowledge acquisition is assessed, and Phase 2, in which knowledge application is assessed. As an illustration, consider the MicroDYN task *handball training* displayed in Figure 1. There, input variables (i.e., different training strategies labeled Strategy A, Strategy B, Strategy C) influence several output variables (i.e., characteristics of the team labeled Motivation, Power of the throw, Exhaustion). In Phase 1, students can freely explore the task (duration: 3 min) by manipulating the sliders on the left and by observing subsequent changes in the output variables on the right (cf. Figure 1). During this free exploration, students are asked to specify the relations between variables on a concept map displayed at the bottom of Figure 1 by drawing arrows between input and output variables (e.g., between Strategy A and Motivation), thereby capturing their mental representation of the underlying system structure. In Phase 2, students are instructed to reach given goal values on the output variables (e.g., increasing the Power of the throw to five) by manipulating the input variables in the correct way (e.g., increasing Strategy B; duration: 1.5 min). Each MicroDYN task is embedded in a different cover story, and inputs as well as outputs are labeled without deep semantic meaning to increase motivation and minimize the influence of prior knowledge.

Depending on the specific number of tasks, a CPS assessment with MicroDYN takes between 40 and 60 min including instructions. Detailed information on the rationale underlying these types of tasks can be found in Funke (2001), and the MicroDYN approach is described in detail in Greiff et al. (2012) and Schweizer et al. (2013).

In all three samples, a set of MicroDYN tasks was used to capture knowledge acquisition and knowledge application as core dimensions of CPS (eight tasks in Study A, 10 tasks in Study B, and seven tasks in Study C). However, with regard to differences in cognitive potential across the three samples (e.g., high-ability university students with above-average cognitive performance in Study A and average-ability high school students in Grade 8 in Study C), MicroDYN task difficulty was adjusted accordingly. To increase difficulty for the more able samples, the underlying system structures of the MicroDYN tasks were designed to be more complex by increasing the number of inputs and outputs, by increasing the number of relations between them, and by introducing outputs that changed by themselves over time without active manipulation of the inputs (for details on altering difficulty in MicroDYN, cf. Greiff et al., 2012).

With regard to the scoring of MicroDYN, full credit for knowledge acquisition was given if students' models contained no mistakes. If additional relations were reported or actual relations were



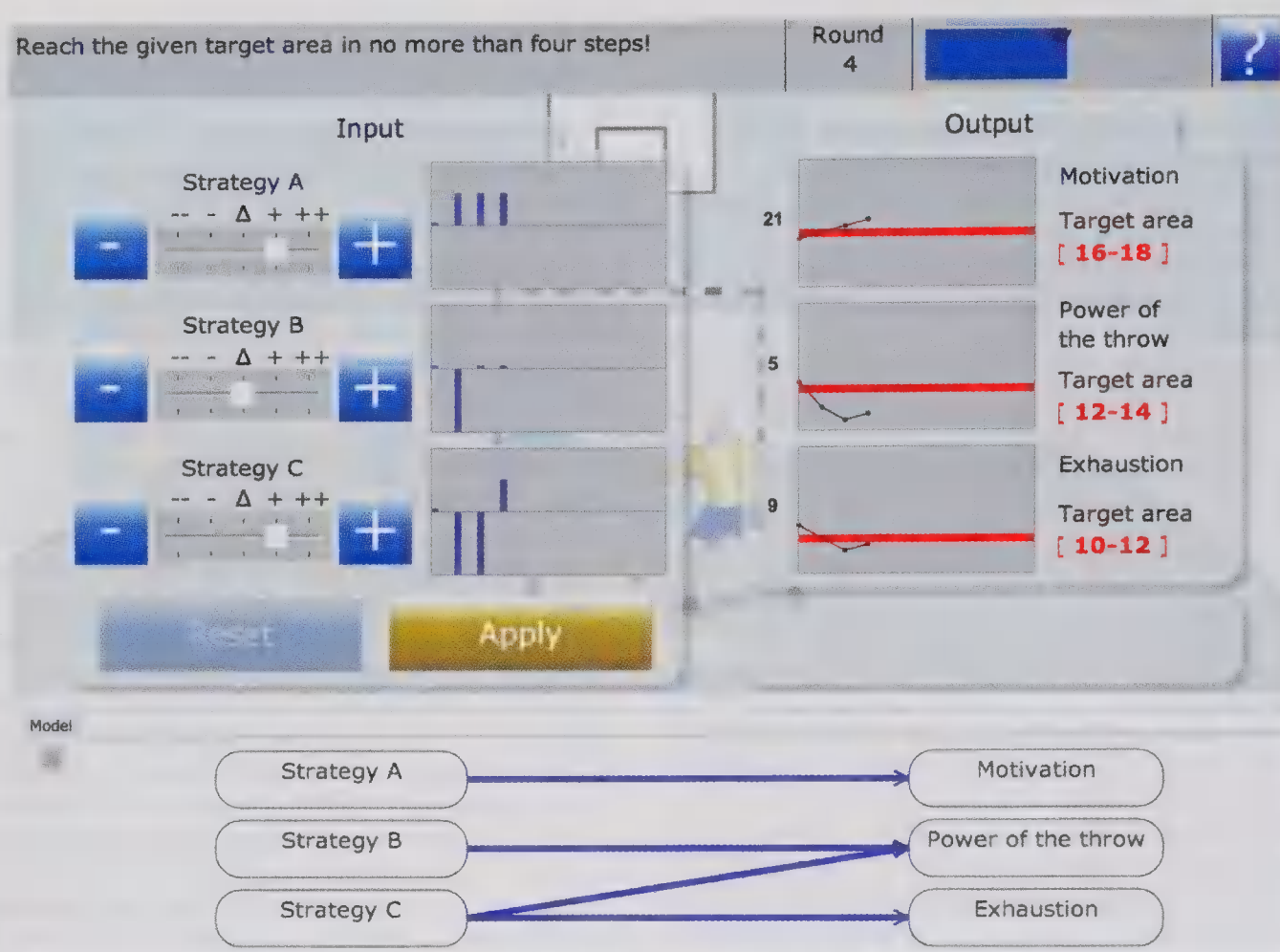


Figure 1. Screenshot of the MicroDYN task *handball training*. The controllers of the input variables range from “- -” (value = -2) to “+ +” (value = +2). The current values and the target values of the output variables are displayed numerically (e.g., current value for Motivation: 21; target values: 16–18) and graphically (current value: dots; target value: red line). The correct model is shown at the bottom of the figure (cf. Wüstenberg et al., 2012).

omitted, zero credit was assigned. A full score in knowledge application was given if goal values were reached, whereas no credit was given if target values were not reached (for details on scoring, cf. Greiff et al., 2012; Kröner, Plass, & Leutner, 2005). Thus, each MicroDYN task yielded indicators on knowledge acquisition and knowledge application totaling eight, 10, and seven indicators in Studies A, B, and C, respectively, for each of the two CPS dimensions.

### Study A: Relations Among CPS Components, Computer Knowledge, Computer Anxiety, Figural Reasoning, and Final Grade-Point Average

**Participants.** The final sample consisted of  $N = 222$  high-ability university students (69% female; age:  $M = 22.8$ ;  $SD = 4.0$ ) majoring mainly in psychology. In psychology, admission depends on final school grade-point average (GPA), and the selection process is highly competitive. As a consequence, psychology students at German universities usually have above-average cognitive performance. Students received partial course credit for participation and an additional obol (€5 [about \$6 U.S.]) for working conscientiously. Missing data that occurred due to software problems or a failure of participants to work conscientiously

led to  $n = 16$  exclusions from the initial sample. The study took place in the Department of Psychology at the University of Heidelberg, Germany.

#### Materials.

**CPS.** MicroDYN with eight different tasks was used for the CPS assessment.

**ICT literacy.** ICT literacy was assessed using two subtests from the German inventory for the assessment of computer literacy, computer-related attitudes, and computer anxiety (Revised Computer Literacy Inventory, INCOBI-R; Richter, Naumann, & Horz, 2010). Both tests were administered on computers. The first subscale, Practical Computer Knowledge (PRACOWI), contains 20 written scenarios of commonly occurring computer problems. For each scenario, one of four presented solutions is correct. The subscale is substantially correlated with measures of computer use and predicts the ability to master everyday computer tasks (Appel, 2012; Naumann, Richter, & Groeben, 2001; Richter, Naumann, & Groeben, 2001; Richter et al., 2010). It distinguishes between computer experts and novices (Naumann et al., 2001) and is best described by a one-dimensional model (Richter et al., 2010). The scale shows good internal consistency (Cronbach's  $\alpha = .83$ ), and the items are one-dimensional according to the Rasch model. Thus,

PRACOWI is a reliable and valid (e.g.,  $r = .60$  with basic computer skills) measure of the ability to deal successfully with everyday computer tasks and problems. It represents the declarative knowledge scope of ICT literacy described by Tsai (2002). The second subscale, Computer Anxiety (COMA), captures worries about the personal use of computers and computer-related anxiety. Computer anxiety is seen as a trait that includes cognitive and affective components (Morris, Davis, & Hutchings, 1981; Richter et al., 2010). The items refer to feelings of anxiety (e.g., "Working with the computer makes me uneasy") as well as to cognitions of concern (e.g., "When working with the computer, I am often afraid of breaking something"). The subscale covers the scope of attitudes toward ICT literacy (cf. Van Braak, 2004). Discriminant and criterion validity ( $r = -.33$  with duration of computer experience) and good reliability (Cronbach's  $\alpha = .82$ ) of the COMA have been shown in several samples (e.g., Appel, 2012; Richter et al., 2010). The subscale consists of eight self-report items that are rated on a 5-point Likert scale (from  $-2 = do\ not\ agree$  to  $2 = agree$ ), with higher values indicating higher anxiety.

**General cognitive ability.** Figural reasoning as a general cognitive ability was assessed using a computer-adapted version of the Advanced Progressive Matrices (APM; Raven, 1958). This test is viewed as a valid indicator of fluid intelligence (Raven, Raven, & Court, 1998) and shows good internal consistency (Cronbach's  $\alpha = .85$ ). Each item was scored dichotomously.

**Academic achievement.** Academic achievement was measured as students' self-reported final school GPA at the end of schooling. As usual in German schools, school marks ranged from 1 (*excellent*) to 6 (*poor*). For further analyses, we reversed the school marks so that higher numerical values reflected better performance.

**Procedure.** Testing was split into two sessions, each lasting approximately 50 min. In the first session, students worked on MicroDYN. In the second session, the APM, PRACOWI, and COMA were administered. Afterwards, students provided demographic data as well as school marks.

### Study B: Relations Among CPS Components, Basic Computer Skills, Computer Anxiety, Figural Reasoning, and Final School Marks

**Participants.** The sample consisted of 341 university students (67% female; age:  $M = 22.3$ ;  $SD = 4.0$ ) with a broad study background who were majoring mainly in social sciences. Students received either partial course credit or a financial reimbursement of €20 (about \$25 U.S.) for their participation. The study took place in the Department of Psychology at the University of Heidelberg, Germany.

#### Materials.

**CPS.** MicroDYN with 10 different tasks was used for CPS assessment.

**ICT literacy.** ICT literacy was assessed with two different instruments. First, a further developed version of the Basic Computer Skills Test (BCS; Goldhammer et al., 2013) was used; it is considered a computer-based, objective, and performance-based measure of basic ICT skills in line with Tsai (2002). The 20 tasks require students to access, collect, and provide information in simulated graphical user interfaces of several computer environments (e.g., web browser, text editor). The environments, although

only an abstract representation of real software, share general characteristics of real computer environments (for more details and task descriptions, see Goldhammer et al., 2013). Further, Goldhammer et al. (2013) reported substantial correlations with other measures of computer skills (e.g.,  $r = .60$  with PRACOWI), discriminant validity (e.g.,  $r = .32$  with word recognition), unidimensionality, and good reliability (Cronbach's  $\alpha = .70$ ). Therefore, the BCS can be considered to be a valid measure of ICT literacy. For each task, the correct user response (BCS ability according to Goldhammer et al., 2013) was given full credit; otherwise, no credit was given. As a second measure, the COMA of the INCOBI-R (Richter et al., 2010) as in Study A (see above) was used.

**General cognitive ability.** Figural reasoning as a general cognitive ability was assessed using a computer-adapted version of the matrices subtest of the Intelligence Structure Test-Revised (Beauducel, Liepmann, Horn, & Brocke, 2010). This test is viewed as a good indicator of reasoning (cf. Carroll, 1993) but consists of more diverse task contents than the APM test used in Study A. According to the test manual, the matrices subtest showed an acceptable reliability (Cronbach's  $\alpha = .71$ ) and validity (Beauducel et al., 2010). Each item of the subtest was scored dichotomously.

**Academic achievement.** Academic achievement was reported as final school marks when leaving high school in four natural science subjects (math, physics, chemistry, and biology) and five subjects that consisted of either social sciences or languages (German, English, history, geography, and social studies). School marks were reversed for all analyses so that higher numerical values reflected better performance.

**Procedure.** Testing was divided into two sessions of 2.5 and 2 hr. In the first session, students worked on MicroDYN and provided demographic data as well as school marks. In the second session, the IST, BCS, and COMA were administered. Additional measures that were not relevant for this article were administered in both the first and second sessions.

### Study C: Relations Among CPS Components, Computer Anxiety, Working Memory Capacity, and Annual School Marks

**Participants.** The sample consisted of 389 high school students (60% female; age:  $M = 17.1$ ;  $SD = 1.1$ ). Students were offered individual feedback on their results in return for their participation. From the initial sample,  $n = 16$  students were excluded from the analyses because of software errors. The study took place at two German high schools, both located in southwestern Germany.

#### Materials.

**CPS.** MicroDYN with seven different tasks was used for the CPS assessment.

**ICT literacy.** ICT literacy was assessed using the COMA from the INCOBI-R (Richter et al., 2010) as in Study A.

**General cognitive ability.** Numerical and spatial working memory capacity were assessed as measures of general cognitive ability using a computer version of the memory updating numerical task (MUN; Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000; Sander, 2005). The aim of the MUN task is to remember the values and the locations of several numbers displayed on the screen, to mentally modify the values according to the task ("up-



dating”), and to return the modified numbers in the proper places (for a detailed description, see Schweizer et al., 2013). Thus, the MUN requires the storage and transformation of information as well as its coordination. The MUN task is used as a marker task for working memory (cf. Oberauer, Süß, Wilhelm, & Wittmann, 2003) because of its good reliability (Cronbach’s  $\alpha = .81$ ) and validity (e.g., substantial factor loadings on the working memory factor simultaneous storage and transformation and the factor coordination). For each item, the percentage of correctly reproduced numbers was used as the performance indicator.

**Academic achievement.** Academic achievement was reflected in school marks from the latest annual school certificate in four natural science subjects (math, physics, chemistry, and biology) and three social science subjects (history, geography, and social studies). School marks were reversed for all analyses so that higher numerical values reflected better performance.

**Procedure.** Testing consisted of one session of 1.5 hr. Students worked on MicroDYN, the MUN, and COMA and provided demographic data as well as school marks. Additional measures that were not relevant for this article were administered subsequently.

## Summary of Measures

Overall, two dimensions of CPS, knowledge acquisition and knowledge application, were assessed on the computer by MicroDYN (in all three studies); ICT literacy was assessed on the computer by measuring practical computer knowledge (PRACOWI; Study A), basic computer skills (BCS; Study B), and computer anxiety (COMA; Studies A, B, and C); general cognitive ability was assessed on the computer by measuring figural reasoning (the APM in Study A, a subtest of the IST in Study B) and working memory capacity (Study C); and academic achievement was measured by overall GPA (Study A), final school marks when leaving high school (Study B), and school marks from the students’ latest annual school certificate (Study C).

## Statistical Methods

Data were analyzed using confirmatory factor analysis (CFA) and structural equation modeling (SEM; cf. Bollen, 1989); that is, all reported results and coefficients were measured on a latent level without measurement error. All models were estimated with the software MPlus 7.0 (Muthén & Muthén, 2012). To evaluate model fit for SEM, we applied standard fit indices such as the comparative fit index (CFI), Tucker-Lewis Index (TLI), root-mean-square error of approximation (RMSEA), standardized root-mean-square residual (SRMR), and weighted root-mean-square residual (WRMR) by endorsing the cutoff values recommended by Hu and Bentler (1999). Unless noted otherwise, we applied standard maximum likelihood estimation and used the full information maximum likelihood estimation method to adjust for missing data in order to ensure high statistical power for the detection of small effects.

In all three studies, we used the following statistical analyses for each research question:

To tackle the first research question, we analyzed the relation between CPS and ICT literacy by computing latent correlations (i.e., correlations adjusted for measurement error).

To tackle the second research question, we analyzed latent relations between CPS and general cognitive ability as criteria and ICT literacy as a predictor in a first model. In this model, the path coefficients from ICT literacy to CPS and general cognitive ability indicated the corresponding impact of ICT literacy. However, even if the path coefficient from ICT literacy to CPS were stronger than the path coefficient from ICT literacy to general cognitive ability, it would not be clear whether ICT literacy had a greater influence on CPS or whether the variation in path coefficients occurred merely by chance. To test this question statistically, we modified the first model to derive a second model. In this second model, the path coefficients from ICT literacy to CPS and general cognitive ability were constrained to equality to simulate an equal impact of ICT literacy on CPS and general cognitive ability. The subsequent change in model fit between the first and the second models provided the answer about whether the difference in impact of ICT literacy was statistically meaningful. If the chi-square difference between the unconstrained (i.e., first) and constrained (i.e., second) model turned out to be significant, this would indicate that constraining the parameters to equality significantly worsened the model fit, and, therefore, we would have to assume an unequal impact of ICT literacy on CPS and general cognitive ability. If several measures of ICT literacy were used in one study (i.e., in Studies A and B), constraints were imposed separately for each measure in order to be able to quantify the impact of the specific ICT literacy measure on CPS and general cognitive ability. We used chi-square difference tests with the mean- and variance-adjusted maximum likelihood estimator (cf. Muthén & Muthén, 2012) for these analyses.

To tackle the third research question, we analyzed the incremental validity of CPS with regard to academic achievement as the criterion in all three studies. In other words, after controlling for general cognitive ability and ICT literacy, we entered CPS as a third predictor in the equation. In detail, we checked the predictive validity of ICT literacy and general cognitive ability in a single model. In a second step, we further added CPS and thus entered all constructs into one model. In this latter model, CPS was regressed on general cognitive ability and ICT literacy first. The CPS residuals of this regression as well as general cognitive ability and ICT literacy were then used to predict academic achievement. If the path coefficients of the CPS residuals ended up being significant, this would indicate that CPS explained additional variance above and beyond general cognitive ability and ICT literacy. That is, the added value of CPS would then not be attributable to an indirect assessment of ICT literacy within CPS measures (for more details on this specific regression procedure, see Wüstenberg et al., 2012). When several indicators of academic achievement were available (Studies B and C), we used CFA to calculate latent grade factors (one factor for natural sciences and one factor for social sciences and languages) instead of using a manifest grade marker of academic achievement (Study A). Furthermore, if the indicators for academic achievement were ordered categorical variables, we used the weighted least squares mean- and variance-adjusted estimator (Muthén & Muthén, 2012) for the statistical analysis of the last research question.

## Results

The purpose of this article was to examine the influence of ICT literacy on CPS. Therefore, instead of turning our attention to verifications of different measurement models, we referred to already existing research to derive the measurement models. That is, the structure and dimensionality as well as the model fit of the measurement models were described in the corresponding articles: for CPS and general cognitive ability, Wüstenberg et al. (2012; Study A), Greiff, Wüstenberg, et al. (2013; Study B), and Schweizer et al., (2013; Study C); for basic computer skills, Goldhammer et al. (2013; Study B); and for practical computer knowledge and computer anxiety, Richter et al. (2010; Study A and all three studies, respectively). On this basis, we created parcels (according to the item-to-construct balance recommended by Little, Cunningham, Shahar, & Widaman, 2002) for each measurement (i.e., measures of CPS, ICT literacy, and general cognitive ability) in order to better capture the latent constructs and to increase the accuracy of parameter estimations. The model fit for all parceled measurement models in our study was at least acceptable (i.e., CFI and TLI  $> .95$ ; RMSEA  $< .06$ ; SRMR  $< .05$  or WRMR  $< .90$ ). Comprehensive correlation tables are available in the supplementary material.

### Results for Research Question 1: How Are ICT Literacy and CPS Related to Each Other?

For each analysis used to address this research question, all structural models showed good model fit (i.e., CFI and TLI  $> .95$ ; RMSEA  $< .06$ ; SRMR  $< .05$  or WRMR  $< .90$ ).

**Study A.** The latent correlation between the two measures of ICT literacy used in Study A (i.e., practical computer knowledge and computer anxiety) of  $r = -.73$  ( $p < .01$ ) was about the same size as the original  $r = -.59$  reported by Richter et al. (2010). However, the latter correlation was on a manifest level and was thus uncorrected for measurement error, whereas the former was corrected for measurement error. Correlations between practical computer knowledge and both knowledge acquisition ( $r = .44$ ,  $p < .01$ ) and knowledge application ( $r = .36$ ,  $p < .01$ ) were moderate in size. Furthermore, correlations between computer anxiety and both knowledge acquisition ( $r = -.25$ ,  $p < .01$ ) and knowledge application ( $r = -.20$ ,  $p < .05$ ) were small in size but statistically significant.

**Study B.** The two measures of ICT literacy used in Study B (i.e., basic computer skills and computer anxiety) were moderately correlated ( $r = -.30$ ,  $p < .01$ ). Correlations between basic computer skills and both knowledge acquisition ( $r = .58$ ,  $p < .01$ ) and knowledge application ( $r = .63$ ,  $p < .01$ ) were large and significant. Smaller, but still significant, were the correlations between computer anxiety and both knowledge acquisition ( $r = -.22$ ,  $p < .01$ ) and knowledge application ( $r = -.31$ ,  $p < .01$ ). Although the latter were slightly higher than in Study A, the relations between computer anxiety and CPS were similar between the two studies.

**Study C.** There were small but significant correlations between computer anxiety and both knowledge acquisition ( $r = -.17$ ,  $p < .01$ ) and knowledge application ( $r = -.21$ ,  $p < .01$ ). Again, the sizes of the coefficients were similar to the other studies.

In conclusion, in all three studies, there were significant relations between different operationalizations of ICT literacy and CPS. In detail, the relations between both knowledge-based and behavioral measures of ICT literacy (i.e., practical computer knowledge and basic computer skills) and CPS were higher than the relation between attitudinal measures of ICT literacy (i.e., computer anxiety) and CPS. The expected modest correlations between the CPS components and operationalizations of ICT literacy indicate that the two constructs are separable.

### Results for Research Question 2: Does ICT Literacy More Strongly Predict a CBA of CPS Than ICT Literacy Predicts the Assessment of General Cognitive Ability?

**Study A.** The model with practical computer knowledge and computer anxiety as simultaneous predictors of CPS and reasoning showed a good overall model fit (Model A.1 in Table 1; see Figure 2 for an illustration of the type of model evaluated in Research Question 2). As expected, practical computer knowledge predicted knowledge acquisition ( $\beta = .59$ ,  $p < .01$ ), knowledge application ( $\beta = .47$ ,  $p < .01$ ), and reasoning ( $\beta = .55$ ,  $p < .01$ ). However, computer anxiety was not a significant predictor at all in this model (knowledge acquisition:  $\beta = -.23$ ; knowledge application:  $\beta = -.20$ ; reasoning:  $\beta = -.19$ ; all  $ps > .05$ ; see Figure 2) even though in the bivariate analysis used to address Research Question 1, computer anxiety was correlated with CPS. The high correlation

Table 1  
Fit Indices for Different Models for Research Question 2

Model	$\chi^2$	<i>df</i>	<i>p</i>	CFI	TLI	RMSEA	WRMR
Model A.1	116.43	94	.06	.976	.969	.037	.798
Model A.2: constrained for PRACOWI	117.43	96	.07	.977	.971	.035	.800
Model A.3: constrained for COMA	117.12	96	.07	.977	.972	.035	.797
Model B.1	90.08	80	.21	.987	.983	.024	.908
Model B.2: constrained for BCS	91.39	82	.22	.988	.984	.023	.907
Model B.3: constrained for COMA	93.49	82	.18	.985	.981	.026	.962
Model C.1	69.55	59	.16	.992	.989	.023	.754
Model C.2: constrained for COMA	77.91	61	.07	.987	.984	.029	1.157

Note.  $\chi^2$  and *df* estimates are based on mean- and variance-adjusted maximum likelihood. CFI = comparative fit index; TLI = Tucker-Lewis Index; RMSEA = root-mean-square error of approximation; WRMR = weighted root-mean-square residual; PRACOWI = Practical Computer Knowledge; COMA = Computer Anxiety; BCS = Basic Computer Skills.



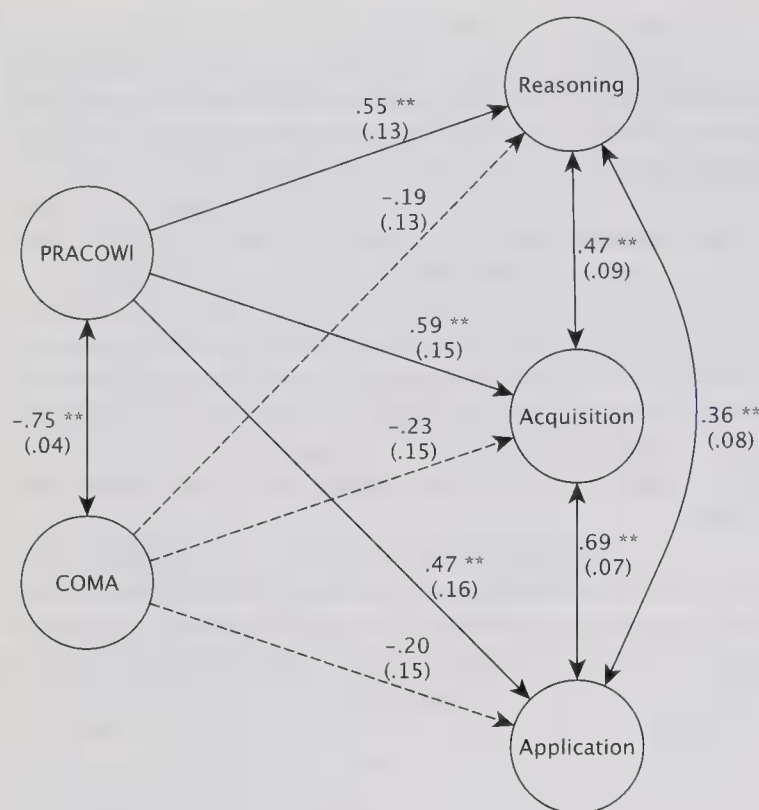


Figure 2. Model A.1 for Research Question 2. Reasoning, knowledge acquisition, and knowledge application were regressed on practical computer knowledge (PRACOWI) and computer anxiety (COMA). Parcels are not depicted. Standard errors are in parentheses.  $^{**}p < .001$ .

between practical computer knowledge and computer anxiety (see Study A with regard to Research Question 1) as well as the lower power of this analysis due to the increased number of variables in the model in Figure 2 were possible explanations for the nonsignificant prediction of computer anxiety.

If the paths from practical computer knowledge to CPS and reasoning were constrained to equality (Model A.2), model fit did not decrease significantly,  $\chi^2(2) = 1.338$ ,  $p > .05$ . Constraining the paths from computer anxiety to CPS and reasoning to equality did not significantly decrease model fit either (Model A.3),  $\chi^2(2) = 0.670$ ,  $p > .05$ . Thus, CPS was not more strongly predicted by ICT literacy than was general cognitive ability.

**Study B.** The model in which basic computer skills and computer anxiety simultaneously predicted CPS and reasoning showed a good overall model fit (Model B.1 in Table 1). Basic computer skills predicted knowledge acquisition ( $\beta = .44$ ,  $p < .01$ ), knowledge application ( $\beta = .52$ ,  $p < .01$ ), and reasoning ( $\beta = .47$ ,  $p < .01$ ). Computer anxiety was a significant predictor of knowledge acquisition ( $\beta = -.14$ ,  $p < .05$ ) and knowledge application ( $\beta = -.23$ ,  $p < .05$ ), but not of reasoning ( $\beta = -.04$ ,  $p > .05$ ). If the paths from basic computer skills to CPS and reasoning were constrained to equality, the model did not fit significantly worse (Model B.2),  $\chi^2(2) = 1.616$ ,  $p > .05$ . Constraining the paths from computer anxiety also did not significantly decrease the model fit (Model B.3),  $\chi^2(2) = 4.177$ ,  $p > .05$ .<sup>1</sup> In sum, CPS was not more strongly predicted by ICT literacy than was general cognitive ability.

**Study C.** The model with computer anxiety as a predictor of CPS and working memory capacity showed a good overall model

fit (Model C.1 in Table 1). Computer anxiety predicted knowledge acquisition ( $\beta = -.18$ ,  $p < .01$ ) and knowledge application ( $\beta = -.24$ ,  $p < .01$ ) but not working memory capacity ( $\beta = -.01$ ,  $p > .05$ ). If the paths from computer anxiety to CPS and working memory capacity were constrained to equality, the model fit decreased significantly (Model C.2),  $\chi^2(2) = 9.027$ ,  $p < .05$ . Results indicated that CPS was more strongly predicted by ICT literacy than was general cognitive ability.

In summary, the findings of two out of the three studies demonstrated that CPS was not more strongly predicted by ICT literacy than was general cognitive ability. In detail, both behavioral and attitudinal operationalizations of ICT literacy impacted CPS in a manner that was similar to their impact on different assessments of figural reasoning. However, CPS was more strongly predicted by attitudinal measures of ICT literacy (i.e., computer anxiety) than was working memory capacity.

### Results for Research Question 3: Can the Added Value of CPS Be Explained by ICT Literacy or Are Distinct Cognitive Processes in CPS Responsible for the Added Value?

**Study A.** In these analyses, we used GPA as a manifest variable. The first model with reasoning, practical computer knowledge, and computer anxiety as predictors of GPA (manifest) showed a good model fit (Model A.1 in Table 2). However, only reasoning ( $\beta = .39$ ,  $p < .01$ ) and computer anxiety ( $\beta = -.25$ ,  $p < .05$ ) significantly predicted GPA but practical computer knowledge did not ( $\beta = .05$ ,  $p > .05$ ). Altogether, about 16% of the variance in GPA was explained in this model. In a second model (A.2), the residuals of CPS after controlling for reasoning and ICT literacy were added simultaneously to the predictors that were already included in the first model (see Figure 3 for an illustration of the type of model evaluated in Research Question 3). Again, GPA was significantly predicted by reasoning ( $\beta = .40$ ,  $p < .01$ ) and computer anxiety ( $\beta = -.25$ ,  $p < .05$ ) but not by practical computer knowledge ( $\beta = .06$ ,  $p > .05$ ). Furthermore, the residuals of CPS after controlling for reasoning, computer anxiety, and practical computer knowledge predicted GPA beyond general cognitive ability and ICT literacy (residuals of knowledge acquisition:  $\beta = .24$ ,  $p < .05$ ; residuals of knowledge application:  $\beta = -.09$ ,  $p > .05$ ; see Figure 3). In the second model, 21% of the variance was explained by CPS, indicating that 5% of the variance was additionally explained in comparison to the first model.

**Study B.** The measurement model for school marks as the criterion in these analyses with the two dimensions natural sciences and social sciences along with languages showed a good model fit,  $\chi^2(26) = 24.03$ ,  $p > .05$ ; CFI = 1.000; TLI = 1.000; RMSEA = .000; WRMR = .528.

Beginning with a model in which final school marks in the natural sciences were significantly predicted by reasoning ( $\beta = .48$ ,  $p < .01$ ) but not by basic computer skills ( $\beta = -.04$ ,  $p > .05$ )

<sup>1</sup> The maximum difference in path size was between knowledge application in CPS and reasoning. Therefore, we tested another more conservative model. However, the results for this more conservative model did not change if just the paths from computer anxiety to knowledge application and reasoning were constrained to equality, whereas the path to knowledge acquisition was freely estimated,  $\chi^2(1) = 3.252$ ,  $p > .05$ .

Table 2  
Fit Indices for Different Models for Research Question 3

Model	$\chi^2$	df	p	CFI	TLI	RMSEA	WRMR
Model A.1: g, PRACOWI, and COMA predict GPA	49.714	39	.12	.990	.987	.035	.035 <sup>a</sup>
Model A.2: g, PRACOWI, COMA, and CPS predict GPA	134.39	105	.03	.983	.977	.036	.041 <sup>a</sup>
Model B.1: g, BCS, and COMA predict school marks	142.33	125	.14	.984	.980	.020	.741
Model B.2: g, BCS, COMA, and CPS predict school marks	260.09	233	.11	.980	.977	.018	.733
Model C.1: MUN and COMA predict school marks	94.33	71	.03	.987	.984	.029	.697
Model C.2: MUN, COMA, and CPS predict school marks	193.98	155	.02	.984	.980	.025	.695

Note.  $\chi^2$  and df estimates are based on maximum likelihood (ML; Models A.1 and A.2) and weighted least squares mean- and variance-adjusted estimator (Models B.1 to C.2 because of ordered categorical school marks), respectively. CFI = comparative fit index; TLI = Tucker-Lewis Index; RMSEA = root-mean-square error of approximation; WRMR = weighted root-mean-square residual; g = Reasoning; PRACOWI = Practical Computer Knowledge; COMA = Computer Anxiety; GPA = grade-point average; CPS = complex problem solving; BCS = Basic Computer Skills; MUN = memory updating numerical.

<sup>a</sup> Standardized root-mean-square residual because of ML estimator for Models A.1 and A.2.

or by computer anxiety ( $\beta = -.01, p > .05$ ), and school marks in the social sciences and languages were predicted only by computer anxiety ( $\beta = -.20, p < .05$ ) but not by reasoning ( $\beta = .17, p > .05$ ) or by basic computer skills ( $\beta = .14, p > .05$ ), we found 21% explained variance in school marks in the natural sciences and 8% explained variance in school marks in the social sciences and languages. The overall model fit was good (see Model B.1 in Table 2). In a second model with good overall model fit (Model B.2 in Table 2) and with the CPS residuals as an additional predictor, there was no substantial change in the pattern of results for reasoning (natural sciences:  $\beta = .48, p < .01$ ; social sciences and languages:  $\beta = .15, p > .05$ ), basic computer skills (natural sciences:  $\beta = -.04, p > .05$ ; social sciences and languages:  $\beta = .15, p > .05$ ), and computer anxiety (natural sciences:  $\beta = .00, p > .05$ ; social sciences and

languages:  $\beta = .20, p < .05$ ). Additionally, the residuals of knowledge application significantly predicted school marks in the natural sciences ( $\beta = .18, p < .05$ ), but no variance in school marks in the social sciences and languages was incrementally predicted by CPS. Overall, 24% of the variance in the natural sciences and 8% of the variance in the social sciences and languages were explained by the second model, indicating that 3% of the variance in the natural sciences and 0% in the social sciences and languages were additionally explained when CPS was included as a third predictor.

**Study C.** The measurement model for school marks as the criterion in these analyses with the two dimensions natural sciences and social sciences showed a good model fit,  $\chi^2(13) = 19.56, p > .05$ ; CFI = .986; TLI = .978; RMSEA = .036; WRMR = .637.

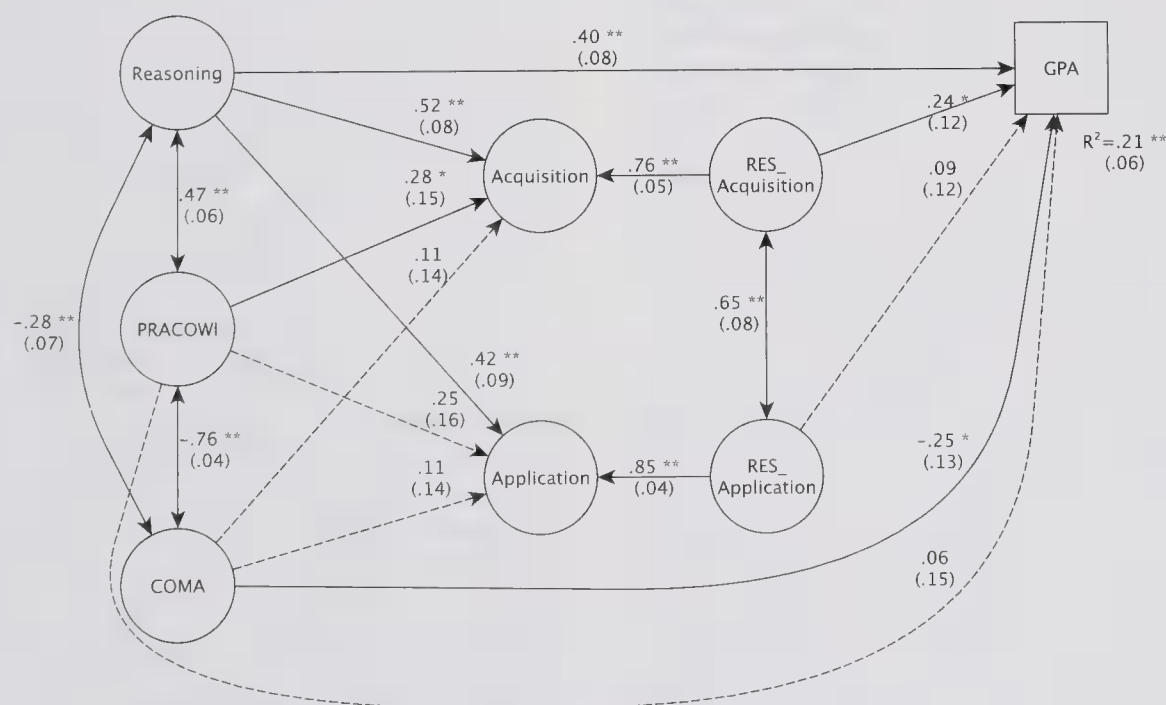


Figure 3. Model A.2 for Research Question 3. Knowledge acquisition and knowledge application were regressed on reasoning, practical computer knowledge (PRACOWI), and computer anxiety (COMA). The computer problem-solving residuals (RES) of this regression as well as reasoning, practical computer knowledge, and computer anxiety were used to predict grade-point average (GPA). Parcels are not depicted. Standard errors are in parentheses. \*  $p < .05$ . \*\*  $p < .001$ .



The first model with working memory capacity and computer anxiety as predictors of annual school marks showed a good model fit (Model C.1 in Table 2). Working memory capacity significantly predicted natural science school marks ( $\beta = .26, p < .01$ ), but computer anxiety did not ( $\beta = -.02, p > .05$ ). Similar results held for the social sciences. There, working memory capacity was a significant predictor ( $\beta = .11, p < .05$ ), but computer anxiety was not ( $\beta = .00, p > .05$ ). For school marks in the natural sciences, 7% of the variance was explained, and for social science school marks, 1% was explained. If CPS residuals were included in the second model (Model C.2) as additional predictors, only the residuals of knowledge acquisition significantly predicted marks in the natural sciences ( $\beta = .25, p < .01$ ) and social sciences ( $\beta = .26, p < .01$ ). There was no change in the pattern of results for working memory capacity (natural sciences:  $\beta = .26, p < .01$ ; social sciences and languages:  $\beta = .11, p < .05$ ) and computer anxiety (natural sciences:  $\beta = -.02, p > .05$ ; social sciences and languages:  $\beta = .00, p > .05$ ). In the second model, 18% of the variance in natural science school marks and 7% of the variance in social science school marks were explained, indicating that 11% and 6%, respectively, were additionally explained by including the CPS residuals in comparison to the first model.

Overall, all three studies demonstrated an added value of CPS beyond different operationalizations of general cognitive ability and ICT literacy. In detail, CPS additionally explained up to 11% of the variance in academic achievement. As indicated by the findings of Studies B and C, CPS was a stronger predictor of academic achievement in the natural sciences than in the social sciences and languages.

## Discussion

The aim of the current study was to deepen the understanding of how individual CPS skills, which are currently receiving considerable interest in educational contexts as a highly relevant transversal skill (Mayer & Wittrock, 2006), are influenced by ICT literacy. To cover the construct comprehensively, this question was pursued in three different samples with a number of different measures of ICT literacy. Furthermore, we controlled for different general cognitive abilities such as reasoning and working memory when relating CPS and ICT literacy. In general, the results of our study supported the assumption that an assessment of CPS allows complex cognitive processes to be captured. These processes are related to ICT literacy and general cognitive ability to a certain extent but not exclusively so. More specifically, CPS skills were weakly to moderately related to ICT literacy (Research Question 1). However, the relations between CPS and different assessments of ICT literacy were (with one exception) just as strong as between general cognitive ability and ICT literacy (Research Question 2). Most importantly, we were able to determine that the added value of CPS recently reported in the literature is not attributable to an indirect assessment of ICT literacy in CPS measures. That is, ICT literacy assessment and CPS assessment were not confounded in such a way that the validity of the latter was threatened. In fact, when controlling for general cognitive ability and ICT literacy, the incremental validity of CPS in predicting relevant external criteria remained substantial (Research Question 3).

In accordance with previous research (Hartig & Klieme, 2005; Süß, 1996), we found a noteworthy relation between CPS and ICT

literacy. We can therefore repeat Mead and Drasgow's (1993) word of caution that the influence of ICT literacy in CBA should not be underestimated and that any computer-delivered measurement instrument should be carefully designed and examined. This may be even more important for assessment instruments that reflect rather complex skills such as CPS or serious games (cf. Michael & Chen, 2006; Russell et al., 2003) that require a somewhat more complex graphical user interface. However, in contrast to recently reported results by Sonnleitner et al. (2013), the added value of CPS independent of ICT literacy was demonstrated consistently in three studies. Sonnleitner et al. (2013) reported an added value of CPS only if the assessment of academic achievement as a criterion was computer based, an idea that indirectly suggests a strong effect of ICT literacy. A reason for the discrepancy between studies could lie in the different operationalizations of CPS. GeneticsLab, the CPS assessment used by Sonnleitner et al. (2013), requires more advanced human-computer interactions than MicroDYN concerning the documentation of acquired knowledge and thus, arguably, an even higher level of ICT literacy. In MicroDYN, students are asked simply to draw arrows between variables on a concept map displayed at the bottom of the screen (see Figure 1), whereas in the GeneticsLab, the concept map is presented in a separate display, and students are asked to draw the relations and to label the strengths of the relations between the variables on a more comprehensive and also more complicated concept map. As a consequence, the GeneticsLab has a longer instruction phase, several user-interface environments for different CPS dimensions, more differentiated knowledge inquiry, and so forth. These features that are characteristic of CPS may increase the validity of the CPS assessment but at the same time may also increase the potential impact of ICT literacy and, thus, the prediction of computer-based external criteria as reported by Sonnleitner et al. (2013). Taking into account the different results concerning the influence of ICT literacy on CPS, we conclude that the impact of ICT literacy depends on the operationalization of CPS and the specific implementation of the assessment. For MicroDYN as a CPS assessment tool, ICT literacy was no threat to its validity. However, this article's purpose, which was to examine the impact of ICT literacy, should be considered for every new operationalization of CPS.

This study was driven by two mutually exclusive explanations for the added value of CPS. It was argued that either (a) CPS assessment captures unique characteristics and complex cognitive processes that are not inherent to general cognitive ability or (b) the assessment of CPS is a confounded assessment of general cognitive ability and ICT literacy. Our findings did not support the second explanation. With regard to the incremental validity of CPS in particular, we came to the conclusion that the assessment of CPS allows researchers to consider complex cognitive processes beyond general cognitive ability (cf. Raven, 2000), indicated by the finding that up to 11% of the variance in academic achievement was additionally explained by CPS beyond the variance explained by general cognitive ability and ICT literacy even though the prediction of social science and language grades was considerably lower. However, the overall result pattern provides important evidence for the validity of CPS in line with recent research (Greiff et al., 2012; Greiff, Wüstenberg, et al., 2013; Wüstenberg et al., 2012).



Furthermore, we discuss two details of our findings more specifically. First, with regard to the different operationalizations of general cognitive ability, we found a stronger influence of the attitudinal component of ICT literacy on CPS than on working memory. In fact, the MUN tasks that were used to assess working memory in Study C were not related to computer anxiety at all. Working memory is supposed to be a general, albeit basic, cognitive ability (cf. Oberauer, Süß, Wilhelm, & Wittmann, 2008), and, thus, its assessment requires just simple human–computer interactions. In contrast to MicroDYN, for which students are asked to use several input and display elements (e.g., sliders, concept maps, diagrams) in different user interfaces, the only computer interaction in the MUN task is indeed just to successively press a number key on the keyboard to interact within a simple and uniform graphical user interface. Thus, we argue that a less complex human–computer interaction will be less influenced by ICT literacy (i.e., at least by the attitudinal component of ICT literacy). Therefore, our findings are rendered even more powerful because, despite increased interactions within the user interface of our CPS assessment (cf. Figure 1), the predictive validity of CPS was not reduced in any of the three studies.

The second detailed result that is worth mentioning is the differential predictive power of the two CPS dimensions: knowledge acquisition and knowledge application. In previous research (e.g., Schweizer et al., 2013; Wüstenberg et al., 2012), knowledge acquisition was the stronger predictor of academic achievement. However, our findings from Study B with regard to Research Question 3 showed a different pattern: Knowledge application was the strongest predictor rather than knowledge acquisition. There may be two different explanations for this result. First, the two dimensions are empirically separable but are highly correlated as found in previous studies (latent correlation around .70–.80). As a consequence, the differences in the relations of the two dimensions to external criteria could be due to a random capitalization on chance with knowledge acquisition being more strongly related to external criteria in some samples and knowledge application in others. Thus, a replication of the finding in Study B should be the next step taken to gain further insights into this issue. The second explanation addresses the demands that the CPS assessment places on students. As mentioned above, the difficulty of the CPS assessment was adjusted with regard to differences in the cognitive potential of the samples. In Study B, the study with the cognitively most able sample, the participants' goal values when their knowledge application was assessed were more complex and interactive compared with in the two other studies. For example, to reach the given target values in knowledge application, more simultaneous inputs were necessary, and multiple targets had to be considered at the same time. Thus, in Study B, we might have captured complex cognitive processes by placing additional demands on the assessment of knowledge application. These additional demands may require processes beyond the complex processes that were already assessed by the knowledge acquisition dimension. Thus, the predictive power of knowledge application surpassed that of the knowledge acquisition dimension. In general, we interpret this finding as an additional potential of the knowledge application dimension, and this has not yet been examined systematically. In conclusion, to increase knowledge about CPS, further research concerning the differential importance of knowledge acquisition and knowledge application is needed to better understand the

differential predictive power of the two CPS dimensions (cf. Sonnleitner, Brunner, Keller, & Martin, 2014).

Finally, some limitations of this article and outlooks for future research should be discussed. As noted above, we used broad operationalizations of ICT literacy and general cognitive ability; thus, the generalizability of our findings was not limited to single-assessment instruments of ICT literacy and general cognitive ability. Specifically, reasoning and working memory are good indicators of general cognitive ability. However, they do not cover the entire range of general cognitive ability, which additionally encompasses attention, long-term memory, processing speed, perception, verbal ability, crystallized intelligence, and so forth. Furthermore, ICT literacy is composed of a variety of aspects. Not all of them were covered in our study. However, our results provided indications of differential effects of ICT literacy on cognitive abilities. For example, computer anxiety as part of ICT literacy had an influence on CPS but not on working memory capacity. Therefore, future research should explore other cognitive abilities and more diverse operationalizations of ICT literacy when relating them to CPS (cf. Wittmann & Süß, 1999). Additionally, only MicroDYN was used as an assessment of CPS. To expand the nomological network, several operationalizations are necessary. Regarding this, the approaches of finite state automata (Buchner & Funke, 1993) or classical microworlds (Funke, 2001) would be worthwhile extensions to MicroDYN. To this end, not only on the level of operationalizations but also with regard to theoretical considerations, additional efforts are needed to more comprehensively understand CPS in the context of cognitive theories such as the theory of situated cognition (Brown, Collins, & Duguid, 1989) or CHC theory (McGrew, 2009). In conclusion, further operationalizations of general cognitive ability and CPS should be considered and theoretical considerations should be made in future research.

We put forward two explanations for the added value of CPS: the additional assessment of complex cognitive processes beyond general cognitive ability and a confounded assessment of general cognitive ability and ICT literacy. However, one could invoke other factors that may account for the added value of CPS, for example, motivation. The issue of motivation and acceptance has been discussed since the beginning of CPS assessment (cf. Kersting, 1998; Sonnleitner et al., 2012; Vollmeyer & Funke, 1999). Gamelike features and attractive graphical setups may enhance motivation, which in turn may explain the added value of CPS. In our studies, we used attitudinal measures of ICT literacy, but we did not include motivational aspects in our research. If we want to be sure that the added value of CPS is mainly based on complex cognitive processes, it will be necessary to extend our research strategy to comprehensively consider other possible explanations such as motivation as well.

We used three different samples for this research: high-ability university students, university students with a broad study background, and high school students. Although these samples covered three relevant groups from the population of students, they are not completely representative of the entire population. With regard to generalizability (Brennan, 1983), we note that our results may be biased by low variability and may be different in other subgroups (e.g., in adults). In this sense, today's students are described as "digital natives" (Prensky, 2001), indicating a generally high level of ICT literacy and, thus, restricted variance. However, the impact



of ICT literacy may not be linear across the entire range of ability. On the one hand, poor ICT literacy (i.e., not being familiar with the operations necessary to handle the computer) may cause severe difficulties, but on the other hand, a high level of ICT literacy may not be helpful in solving CPS tasks. It is important to consider such a nonlinear relation (cf. Leutner, 2002) for a deeper understanding of the differentiated impact of ICT literacy on CPS. However, such analyses need more heterogeneous samples, which should be considered in future research.

In conclusion, our aim was to extend the understanding of the assessment of CPS and, thus, to make a contribution to the validation of the CBA of transversal skills. CPS as one of the most promising of these skills is an important part of international educational large-scale assessments such as PISA. The results of these large-scale assessments have extensive and substantial implications, for example, on further developments of educational systems. Thus, we believe that our research will lead to a better understanding of the results of educational large-scale assessments and, hopefully, to well-founded decisions that are aimed at benefiting students' transversal skills in a quickly changing world.

## References

- Appel, M. (2012). Are heavy users of computer games and social media more computer literate? *Computers & Education*, 59, 1339–1349. doi:10.1016/j.compedu.2012.06.004
- Autor, D., & Dorn, D. (2009). This job is 'getting old': Measuring changes in job opportunities using occupational age structure. *American Economic Review*, 99, 45–51. doi:10.1257/aer.99.2.45
- Autor, D., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical investigation. *The Quarterly Journal of Economics*, 118, 1279–1333. doi:10.1162/003355303322552801
- Ballantine, J. A., McCourt Larres, P., & Oyeler, P. (2007). Computer usage and the validity of self-assessed computer competence among first-year business students. *Computers & Education*, 49, 976–990. doi:10.1016/j.compedu.2005.12.001
- Beauducel, A., Liepmann, D., Horn, S., & Brocke, B. (2010). *Intelligence Structure Test*. Oxford, England: Hogrefe.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.
- Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing.
- Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, 18, 32–42. doi:10.3102/0013189X018001032
- Buchner, A., & Funke, J. (1993). Finite-state automata: Dynamic task environments in problem-solving research. *The Quarterly Journal of Experimental Psychology Section A*, 46, 83–118. doi:10.1080/14640749308401068
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511571312
- Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology*, 33, 593–602. doi:10.1111/1467-8535.00294
- Danner, D., Hagemann, D., Schankin, A., Hager, M., & Funke, J. (2011). Beyond IQ: A latent state-trait analysis of general intelligence, dynamic decision making, and implicit learning. *Intelligence*, 39, 323–334. doi:10.1016/j.intell.2011.06.004
- Dörner, D., Kreuzig, H. W., Reither, F., & Stäudel, T. (1983). *Lohhausen: Vom Umgang mit Unbestimmtheit und Komplexität* [Lohhausen: On dealing with uncertainty and complexity]. Stuttgart, Germany: Huber.
- Fischer, A., Greiff, S., & Funke, J. (2012). The process of solving complex problems. *Journal of Problem Solving*, 4, 19–41.
- Frensch, P. A., & Funke, J. (1995). *Complex problem solving: The European perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking & Reasoning*, 7, 69–89. doi:10.1080/13546780042000046
- Funke, J. (2010). Complex problem solving: A case for complex cognition? *Cognitive Processing*, 11, 133–142. doi:10.1007/s10339-009-0345-0
- Goldhammer, F., Naumann, J., & Keßel, Y. (2013). Assessing individual differences in basic computer skills. *European Journal of Psychological Assessment*, 29, 236–275. doi:10.1027/1015-5759/a000153
- Gonzalez, C., Thomas, R. P., & Vanyukov, P. (2005). The relationships between cognitive ability and dynamic decision making. *Intelligence*, 33, 169–186. doi:10.1016/j.intell.2004.10.002
- Greiff, S., Fischer, A., Wüstenberg, S., Sonnleitner, P., Brunner, M., & Martin, R. (2013). A multitrait-multimethod study of assessment instruments for complex problem solving. *Intelligence*, 41, 579–596. doi:10.1016/j.intell.2013.07.012
- Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement*, 36, 189–213. doi:10.1177/0146621612439620
- Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts—Something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology*, 105, 364–379. doi:10.1037/a0031856
- Hartig, J., & Klieme, E. (2005). Die Bedeutung schulischer Bildung und soziobiographischer Merkmale für die Problemlösekompetenz [The importance of school education and sociobiographical characteristics for problem solving skills]. In E. Klieme, D. Leutner, & J. Wirth (Eds.), *Problemlösekompetenz von Schülerinnen und Schülern. Diagnostische Ansätze, theoretische Grundlagen und empirische Befunde der deutschen PISA-2000-Studie* (pp. 83–97). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55. doi:10.1080/10705519909540118
- Kersting, M. (1998). Differentielle Aspekte der sozialen Akzeptanz von Intelligenztests und Problemlöseszenarien als Personalauswahlverfahren [Differential-psychological aspects of applicants' acceptance of intelligence tests and problem solving scenarios as diagnostic tools for personnel selection]. *Zeitschrift für Arbeits- und Organisationspsychologie*, 42, 61–75.
- Kröner, S., Plass, J., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence*, 33, 347–368. doi:10.1016/j.intell.2005.03.002
- Kuhn, D. (2009). Do students need to be taught how to reason? *Educational Research Review*, 4, 1–6. doi:10.1016/j.edurev.2008.11.001
- Kyllonen, P. C. (2009). New constructs, methods, and directions for computer-based assessment. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment* (pp. 151–156). Luxembourg, Luxembourg: Office for Official Publications of the European Communities.
- Leutner, D. (2002). The fuzzy relationship of intelligence and problem solving in computer simulations. *Computers in Human Behavior*, 18, 685–697. doi:10.1016/S0747-5632(02)00024-9
- Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, 9, 151–173. doi:10.1207/S15328007SEM0902\_1
- Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 287–303). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

- McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, 37, 1–10. doi:10.1016/j.intell.2008.08.004
- Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114, 449–458. doi:10.1037/0033-2909.114.3.449
- Michael, D. R., & Chen, S. L. (2006). *Serious games: Games that educate, train, and inform*. Boston, MA: Thomson Course Technology.
- Molnár, G., Greiff, S., & Csapó, B. (2013). Inductive reasoning, domain specific and complex problem solving: Relations and development. *Thinking Skills and Creativity*, 9, 35–45. doi:10.1016/j.tsc.2013.03.002
- Morris, L. W., Davis, M. A., & Hutchings, C. H. (1981). Cognitive and emotional components of anxiety: Literature review and a revised worry-emotionality scale. *Journal of Educational Psychology*, 73, 541–555. doi:10.1037/0022-0663.73.4.541
- Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- Naumann, J., Richter, T., & Groeben, N. (2001). Validierung des INCOBI anhand eines Vergleichs von Anwendungsexperten und Anwendungsnovizen [Validation of the INCOBI through comparison of expert and novice computer users]. *Zeitschrift für Pädagogische Psychologie*, 15, 219–232. doi:10.1024/1010-0652.15.34.219
- Novick, L. R., & Bassok, M. (2005). Problem solving. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 321–349). Cambridge, England: Cambridge University Press.
- Oberauer, K., Süß, H.-M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity—facets of a cognitive ability construct. *Personality and Individual Differences*, 29, 1017–1045. doi:10.1016/S0191-8869(99)00251-2
- Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittmann, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence*, 31, 167–193. doi:10.1016/S0160-2896(02)00115-0
- Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittmann, W. W. (2008). Which working memory functions predict intelligence? *Intelligence*, 36, 641–652. doi:10.1016/j.intell.2008.01.007
- Organisation for Economic Co-Operation and Development. (2007). *PISA 2006 science competencies for tomorrow's world*. Paris, France: Author.
- Organisation for Economic Co-Operation and Development. (2009a). Chapter 1: Programme for International Student Assessment: An overview. Retrieved from <http://www.oecd.org/berlin/42174841.pdf>
- Organisation for Economic Co-Operation and Development. (2009b). PIAAC problem solving in technology-rich environments. Retrieved from <http://search.oecd.org/officialdocuments/displaydocumentpdf/?doclanguage=en&cote=edu/wkp%282009%2915>
- Organisation for Economic Co-Operation and Development. (2010). PISA 2012 field trial problem solving framework. Retrieved from <http://www.oecd.org/pisa/pisaproducts/46962005.pdf>
- Organisation for Economic Co-Operation and Development. (2012). Draft PISA 2015 collaborative problem solving assessment framework. Retrieved from <http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf>
- Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York, NY: Springer.
- Prensky, M. (2001). Digital natives, digital immigrants. Part I. *On the Horizon*, 9, 1–6. doi:10.1108/10748120110424816
- Raven, J. (1958). *Advanced progressive matrices* (2nd ed.). London, England: Lewis.
- Raven, J. (2000). Psychometrics, cognitive ability, and occupational performance. *Review of Psychology*, 7, 51–74.
- Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's progressive matrices and vocabulary scales: Section 4. The advanced progressive matrices*. San Antonio, TX: Harcourt Assessment.
- Richter, T., Naumann, J., & Groeben, N. (2001). Das Inventar zur Computerbildung (INCOBI): Ein Instrument zur Erfassung von Computer Literacy und computerbezogenen Einstellungen bei Studierenden der Geistes- und Sozialwissenschaften [The Computer Literacy Inventory - An instrument for the assessment of computer literacy and attitudes toward the computer in students of the humanities and social sciences]. *Psychologie in Erziehung und Unterricht*, 48, 1–13.
- Richter, T., Naumann, J., & Horz, H. (2010). Eine revidierte Fassung des Inventars zur Computerbildung (INCOBI-R) [A revised version of the Computer Literacy Inventory]. *Zeitschrift für Pädagogische Psychologie*, 24, 23–37. doi:10.1024/1010-0652/a000002
- Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-based testing and validity: A look back into the future. *Assessment in Education: Principles, Policy & Practice*, 10, 279–293. doi:10.1080/0969594032000148145
- Sander, N. (2005). *Inhibitory and executive functions in cognitive psychology: An individual differences approach examining structure and overlap with working memory capacity and intelligence*. Aachen, Germany: Shaker.
- Scheuermann, F., & Björnsson, J. (2009). *The transition to computer-based assessment*. Luxembourg, Luxembourg: Office for Official Publications of the European Communities.
- Schweizer, F., Wüstenberg, S., & Greiff, S. (2013). Validity of the MicroDYN approach: Complex problem solving predicts school grades beyond working memory capacity. *Learning and Individual Differences*, 24, 42–52. doi:10.1016/j.lindif.2012.12.011
- Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., . . . Latour, T. (2012). The Genetics Lab: Acceptance and psychometric characteristics of a computer-based microworld assessing complex problem solving. *Psychological Test and Assessment Modeling*, 54, 54–72.
- Sonnleitner, P., Brunner, M., Keller, U., & Martin, R. (2014). Differential relations between facets of complex problem solving and students' immigration background. *Journal of Educational Psychology*, 106, 681–695. doi:10.1037/a0035506
- Sonnleitner, P., Keller, U., Martin, R., & Brunner, M. (2013). Students' complex problem-solving abilities: Their structure and relations to reasoning ability and educational success. *Intelligence*, 41, 289–305. doi:10.1016/j.intell.2013.05.002
- Spitz-Oener, A. (2006). Technical change, job tasks and rising educational demands: Looking outside the wage structure. *Journal of Labor Economics*, 24, 235–270. doi:10.1086/499972
- Süß, H.-M. (1996). *Intelligenz, Wissen und Problemlösen: Kognitive Voraussetzungen für erfolgreiches Handeln bei computersimulierten Problemen [Intelligence, knowledge, and problem solving: Cognitive prerequisites of successful performance in computer-simulated problems]*. Göttingen, Germany: Hogrefe.
- Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 19–30). Cambridge, MA: Cambridge University Press. doi:10.1017/CBO9780511816819.003
- Tsai, M.-J. (2002). Do male and female students often perform better than female students when learning computers? A study of Taiwanese eight graders' computer education through strategic and cooperative learning. *Journal of Educational Computing Research*, 26, 67–85. doi:10.2190/9JW6-VVIP-FAX8-CGE0
- Van Braak, J. P. (2004). Domains and determinants of university students' self-perceived computer competence. *Computers & Education*, 43, 299–312. doi:10.1016/j.compedu.2003.09.006
- Van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing*. Dordrecht, the Netherlands: Kluwer Academic Publishers.
- Vollmeyer, R., & Funke, J. (1999). Personen- und Aufgabenmerkmale beim komplexen Problemlösen [Person and task effects within complex problem solving]. *Psychologische Rundschau*, 50, 213–219. doi:10.1026/0033-3042.50.4.213



- Wenke, D., Frensch, P. A., & Funke, J. (2005). CPS and intelligence: Empirical relation and causal direction. In R. J. Sternberg & J. E. Pretz (Eds.), *Cognition and intelligence: Identifying the mechanisms of the mind* (pp. 160–187). New York, NY: Cambridge University Press.
- Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Wirth, J., & Klieme, E. (2003). Computer-based assessment of problem solving competence. *Assessment in Education: Principles, Policy & Practice*, 10, 329–345. doi:10.1080/0969594032000148172
- Wittmann, W. W., & Süß, H.-M. (1999). Investigating the paths between working memory, intelligence, knowledge, and complex problem-

- solving performances via Brunswik symmetry. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, content determinants* (pp. 77–108). Washington, DC: American Psychological Association. doi:10.1037/10315-004
- Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving—More than reasoning? *Intelligence*, 40, 1–14. doi:10.1016/j.intell.2011.11.003

Received March 1, 2013

Revision received November 11, 2013

Accepted November 17, 2013 ■

## ORDER FORM

Start my 2014 subscription to the *Journal of Educational Psychology*® ISSN: 0022-0663

_____ \$89.00	APA MEMBER/AFFILIATE	_____
_____ \$208.00	INDIVIDUAL NONMEMBER	_____
_____ \$751.00	INSTITUTION	_____
In DC and MD add 6% sales tax		
TOTAL AMOUNT DUE		\$ _____

**Subscription orders must be prepaid.** Subscriptions are on a calendar year basis only. Allow 4–6 weeks for delivery of the first issue. Call for international subscription rates.



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

### SEND THIS ORDER FORM TO

American Psychological Association  
Subscriptions  
750 First Street, NE  
Washington, DC 20002-4242

Call **800-374-2721** or 202-336-5600  
Fax **202-336-5568** :TDD/TTY **202-336-6123**  
For subscription information,  
e-mail: **subscriptions@apa.org**

☐ **Check enclosed** (make payable to APA)

**Charge my:** ☐ Visa ☐ MasterCard ☐ American Express

Cardholder Name \_\_\_\_\_

Card No. \_\_\_\_\_ Exp. Date \_\_\_\_\_

\_\_\_\_\_  
Signature (Required for Charge)

### Billing Address

Street \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

Daytime Phone \_\_\_\_\_

E-mail \_\_\_\_\_

### Mail To

Name \_\_\_\_\_

Address \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

APA Member # \_\_\_\_\_

EDUA14

# Differential Relations Between Facets of Complex Problem Solving and Students' Immigration Background

Philipp Sonnleitner  
University of Luxembourg

Martin Brunner  
Free University of Berlin

Ulrich Keller and Romain Martin  
University of Luxembourg

Whereas the assessment of complex problem solving (CPS) has received increasing attention in the context of international large-scale assessments, its fairness in regard to students' cultural background has gone largely unexplored. On the basis of a student sample of 9th-graders ( $N = 299$ ), including a representative number of immigrant students ( $N = 127$ ), the present study evaluated (a) whether CPS can be assessed fairly among students with or without immigration background and (b) whether achievement differences between these groups exist. Results showed that fair assessment of CPS is possible using the Genetics Lab, a computer-based microworld that incorporates game-like characteristics and multilingual-friendly features. Immigrant students were generally outperformed by their nonimmigrant peers, but performance differences can largely be explained by differential enrollment in lower academic tracks. Interestingly, CPS scales were less affected by students' educational background than a traditional paper-pencil-based reasoning scale. Moreover, a fine-grained analysis of different facets of CPS showed that irrespective of the academic track, immigrant students demonstrated a more efficient task exploration behavior than their native peers ( $d = 0.26$ ). In sum, this might point to the potential of computer-based assessment of CPS to identify otherwise hidden cognitive potential in immigrant students.

**Keywords:** complex problem solving, measurement invariance, students with immigration background, students' exploration behavior, Genetics Lab

**Supplemental materials:** <http://dx.doi.org/10.1037/a0035506.supp>

Educational systems currently face two major challenges. First, pressure has risen to include and assess cross-curricular competencies within new educational curricula (Elliot Bennett, Jenkins, Persky, & Weiss, 2003; Kuhn, 2009; Ridgway & McCusker, 2003). The computer-based assessment of students' complex problem-solving skill (CPS) has been suggested as a possible route to addressing this alluring but diffuse set of abilities (Greiff, Kretzschmar, Müller, Spinath, & Martin, 2014; Greiff et al., 2013). CPS describes the competency to adequately interact with domain-

general problems in order to gather and successfully apply knowledge to reach certain target states (e.g., Buchner, 1995). This applied, domain-general character implies that cognitive processes related to CPS can be used in very different content areas and thus makes CPS a central example of cross-curricular competencies.

Initial results concerning a psychometrically sound and reliable assessment of CPS within the educational context have been promising (Greiff et al., 2013; Sonnleitner et al., 2012). A huge step in this direction has also been taken through the inclusion of CPS measures in one of the most significant international large-scale assessments, the Program for International Student Assessment (PISA; Leutner, Fleischer, Wirth, Greiff, & Funke, 2012; OECD, 2010). However, since the development of CPS assessment instruments relies on fairly recent advances in computer-based assessment, research on this topic is still relatively scarce.

A second challenge facing educational systems today is finding appropriate ways to respond to the specific sociocultural and socioeconomic needs of increasing numbers of students with immigration background (Meunier, 2011; OECD, 2012). The Organisation for Economic Co-operation and Development (OECD) has argued that only when a country's educational system succeeds in adequately integrating immigrant students can these students fully develop their potential to participate in a society's social and economic life (OECD, 2012). Yet most immigrant students lag behind their nonimmigrant peers in various academic subjects (e.g., mathematics; OECD, 2012; Schleicher, 2006). A possible

---

This article was published Online First February 17, 2014.

Philipp Sonnleitner, Centre for Educational Measurement and Applied Cognitive Science, University of Luxembourg, Luxembourg, Luxembourg; Martin Brunner, Berlin-Brandenburg Institute for School Quality Improvement, Free University of Berlin, Berlin, Germany; Ulrich Keller and Romain Martin, Centre for Educational Measurement and Applied Cognitive Science, University of Luxembourg, Luxembourg, Luxembourg.

This work was funded by the National Research Fund Luxembourg (FNR/C08/LM/06). We thank all the students and teachers for participating in this study. A special thank you goes to Carrie Kovacs for her editorial support.

Correspondence concerning this article should be addressed to Philipp Sonnleitner, Centre for Educational Measurement and Applied Cognitive Science (EMACS), University of Luxembourg, Campus Kirchberg, 6, rue Richard Coudenhove-Kalergi, 1359 Luxembourg, Luxembourg. E-mail: philipp.sonnleitner@uni.lu



reason can be seen in the specific challenges they face, such as being educated in a language different to their mother tongue (OECD, 2012).

Consequently, if the teaching and assessment of cross-curricular competencies such as CPS is to be established as an official part of school curricula, the consideration of immigrant students is vital. So far, nothing is known about whether CPS measures are fair with respect to immigration background or whether performance differences exist between immigrant students and their native peers.

The present study attempts to fill this gap by thoroughly exploring whether immigrant students are disadvantaged on measures of CPS or not. Specifically, we investigated whether an established measure of CPS (i.e., the Genetics Lab; Sonnleitner et al., 2012) is measurement invariant and thus fair with respect to immigration background. Only measurement invariance ensures (a) that the test works equally for students with and without immigration background, (b) that the same construct is measured in both groups, and (c) that any performance differences between these groups are due to actual differences in the construct being measured, and not to bias or error produced, for instance, by construct-irrelevant cultural differences (Little, 1997; Widaman & Reise, 1997). In a consecutive step, we attempted to determine whether and to what extent performance differences in individual facets of CPS exist for students of varying immigration backgrounds.

To this end, we drew on a sample of ninth grade students of differing immigration backgrounds enrolled in different academic tracks. The study took place in Luxembourg, a country known for its high ratio of immigrant students (Burton & Martin, 2008; OECD, 2012). This heterogeneous sample may make results particularly relevant on an international level, since high immigrant rates can be found in countries around the globe (OECD, 2012).

In sum, the present study addresses the gap in our understanding of immigrant students' CPS performance by investigating (a) fairness of CPS assessment with regard to immigration background and (b) performance differences between immigrant and native students in individual facets of CPS.

### Complex Problem Solving: Computer-Based Assessment and Performance Scores

Students' skill to solve complex problems is typically assessed by computer-based microworlds, such as the Genetics Lab (Sonnleitner et al., 2012; Sonnleitner, Keller, Martin, & Brunner, 2013) shown in Figure 1. Crucially, such microworlds (a) incorporate several characteristics that also describe problems of high complexity in everyday life and (b) require the students to acquire knowledge about the problem in order to purposefully apply it to reach certain target states of the problem (cf. Funke, 2010).

For better illustration, the Genetics Lab is depicted in Figure 1. In the first, *knowledge acquisition* phase (Figure 1a), students are asked to imagine that they are researchers in a genetics lab where they can manipulate several genes of fictitious creatures in order to study how these genes are related to several characteristics of the creatures. Some of these characteristics additionally change on their own, that is as a function of time. Students can depict the knowledge they have gathered about the relations between genes and creature characteristics in a creature-specific database (Figure 1b) by means of a causal diagram. In the second, *knowledge application* phase (Figure 1c), students must apply the gathered

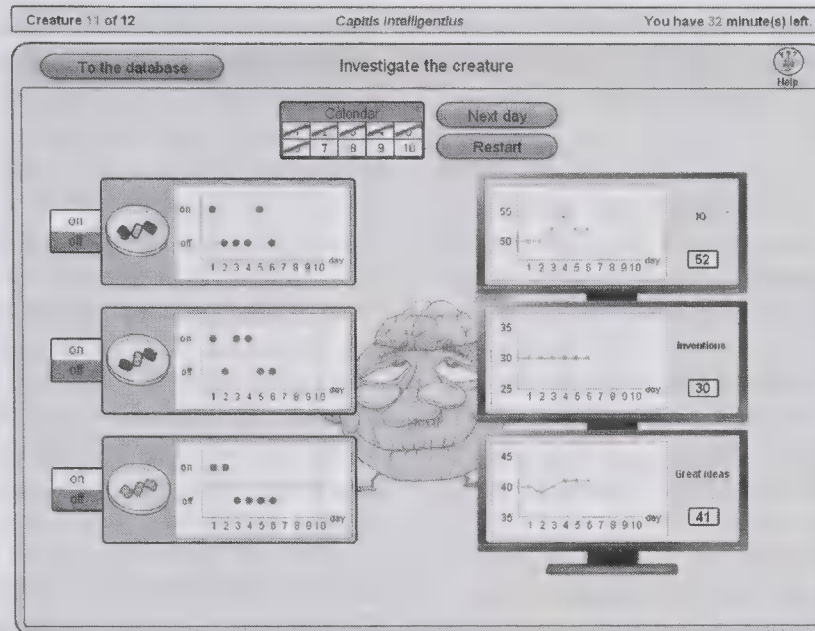
knowledge in order to achieve several target states of the characteristics within a given number of manipulations. Note that the semantic embedding is entirely fictive, making only very low demands on previous knowledge (Greiff, Wüstenberg, & Funke, 2012).

Typically, the Genetics Lab provides three scores, which reflect different facets of students' complex problem solving behavior. The first, *rule identification*, describes the quality and efficiency with which students explore the given problem. Some exploration strategies are more informative than others. For example, it is more informative to manipulate only one gene and then study this manipulation's effects on the creature's characteristics than to manipulate several genes at the same time. Simultaneously manipulating multiple genes means that their individual effects are intermingled and can no longer be unambiguously identified (Vollmeyer, Burns, & Holyoak, 1996). The second score reflects students' skill to express their gathered *rule knowledge*<sup>1</sup> within a causal diagram. Compared to other microworlds assessing CPS (e.g., MicroDYN; Greiff et al., 2012; Wüstenberg, Greiff, & Funke, 2012), the Genetics Lab allows for a more differentiated assessment of *rule knowledge*. Students not only have to show relational knowledge by indicating whether a causal relation exists between a given gene and a certain characteristic; they also have to demonstrate knowledge about the type (increasing or decreasing) and strength (weak or strong) of this relation (Blech & Funke, 2005). The third score, *rule application* (see footnote 1) relates to the students' skill to utilize the gathered knowledge in order to achieve the given target values on the creatures' characteristics. Since the number of available manipulations is limited, students' skill to plan, make forecasts, and react to unexpected consequences comes into play (Funke, 2003).

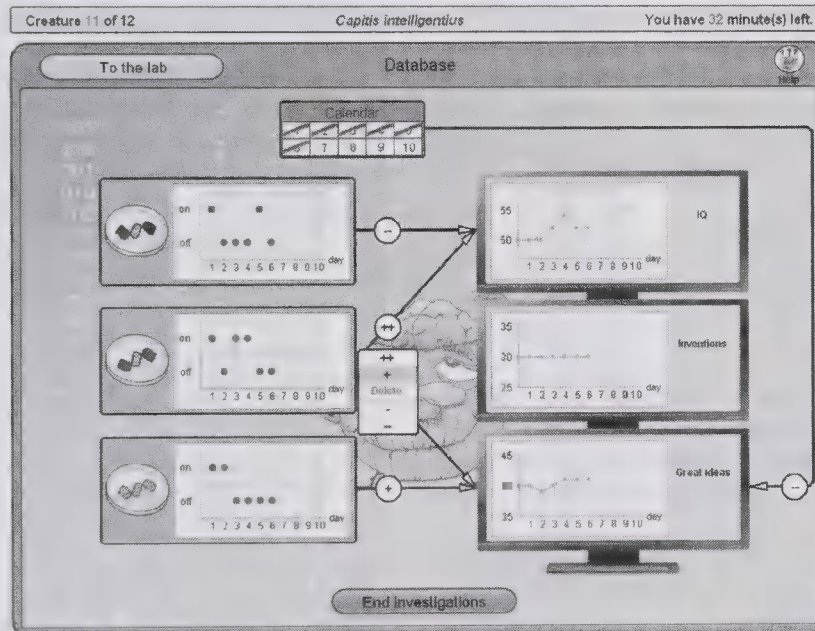
Currently, there is no consensus on how to best represent the various phases of the problem solving process psychometrically. Several researchers, for example, consider CPS to be a multidimensional construct. Thus, they distinguish facets corresponding to all or a subset of the phases and derive scores for *rule identification*, *rule knowledge*, and *rule application*. Interestingly, many studies in the educational domain that have drawn on student samples have only obtained scores for the facets of *rule knowledge* and *rule application* (Bühner, Kröner, & Ziegler, 2008; Greiff et al., 2013; Wüstenberg, Greiff, Molnár, & Funke, 2014). When the third facet of *rule identification* has been measured, its independence from the other two facets has been thrown into question. Whereas Kröner, Plass, and Leutner (2005), as well as Sonnleitner et al. (2013) could reliably measure and discriminate between all three facets of CPS, Schweizer, Wüstenberg, and Greiff (2013) found *rule identification* to be identical to *rule knowledge* and thus redundant. Irrespective of whether two or three facets of CPS were distinguished, all previous studies have shown that the facets of CPS are strongly interrelated. The better students' skill to explore a problem, the higher their acquired knowledge, and the better their skill to reach given target values is. Further, more acquired knowledge is linked to a better skill to reach given target values.

<sup>1</sup> Please note that in several studies that do not assess *rule identification*, the scores *rule knowledge* and *rule application* are labeled according to the related problem solving phases, knowledge acquisition and knowledge application.

### A) Rule Identification



### B) Rule Knowledge



### C) Rule Application

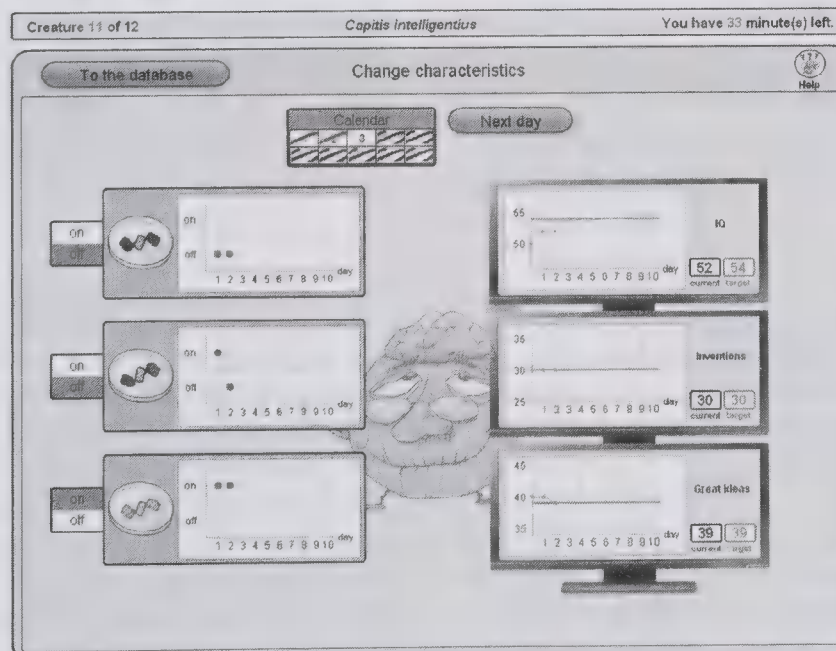


Figure 1 (opposite)



Correlations (adjusted for measurement error) between the facets have typically been well above .60 (Greiff et al., 2013; Kröner et al., 2005; Schweizer et al., 2013; Sonnleitner et al., 2013; Wüstenberg et al., 2012).

Whereas those researchers who consider CPS to be a multidimensional construct derive only facet-specific scores, other researchers are solely interested in a global CPS score (e.g., PISA 2012; Abele et al., 2012; OECD, 2010). In particular, these latter researchers conceive of CPS as a hierarchical construct where a general CPS skill explains the intercorrelations among the facets (Sonnleitner et al., 2013). In this theoretical framework, a higher general CPS skill leads to better performance in all phases of the problem solving process. Since the question of which psychometric conceptualization (multidimensional vs. hierarchical) represents CPS best and provides the most value for applied settings is still open to debate (cf. Sonnleitner et al., 2013), the current study draws on both conceptualizations.

### Measurement Invariance and Group Differences in CPS

The central prerequisite for a fair comparison of performance differences in groups is that the administered measurement instrument essentially measures the same construct in both groups and thus is measurement invariant or measurement equivalent (see Little, 1997, or Widaman & Reise, 1997). According to Little (1997, p. 56),

Measurement equivalence (strong factorial invariance) indicates that (a) the constructs are generalizable to each sociocultural context, (b) sources of bias and error (e.g., cultural bias, translation errors, varying conditions of administration) are minimal, (c) cultural differences have not differentially affected the constructs underlying measurement characteristics [...], and (d) between-culture differences in the constructs' mean, variance, and covariance relations are quantitative in nature (i.e., the nature of cultural differences can be assessed as mean-level, variance, and covariance or correlational effects).

Consequently, a comparison of manifest test scores or latent means across groups is justified and fair only if strong factorial measurement invariance holds. Whether measurement invariance is tenable for a certain measure is usually explored in a stepwise procedure, increasingly constraining model parameters to be the same across groups and investigating whether this significantly impacts model fit (cf. Little, 1997; Widaman & Reise, 1997).

The investigation of measurement invariance regarding CPS is still at its beginning. Importantly, no studies have yet investigated

measurement invariance for all facets of CPS or general CPS for students of differing immigration backgrounds. Nevertheless, there are some studies that provide promising results. The assessment of *rule knowledge* and *rule application* has been found to be measurement invariant and thus fair across sex, nationalities, and grade-levels (Greiff et al., 2013; Wüstenberg et al., 2014). For the facet of *rule identification*, however, measurement invariance has yet to be established.

Group comparisons (i.e., mean comparisons after measurement invariance has been established) of *rule knowledge* and *rule application* have shown a strong influence of educational background (Greiff et al., 2013; Wüstenberg et al., 2014). In general, higher grade level and highest attended educational level corresponded with better performance in *rule knowledge* and *rule application*. In a cross-cultural study reported by Wüstenberg et al. (2014), Hungarian high school students were slightly outperformed by their German counterparts in *rule knowledge* and *rule application*.

### Reasons for Expected Performance Differences in CPS Skill Due to Students' Immigration Background

Despite positive attitudes toward learning and school, immigrant students perform significantly worse than their nonimmigrant peers on mathematics, reading, science and (paper-pencil based measures of) problem solving skills as assessed in PISA 2003 (Martin, Liem, Mok, & Xu, 2012; Schleicher, 2006). There are several additional reasons why performance differences in CPS skill might be expected between immigrant students and their nonimmigrant peers. First, especially for the facet of *rule identification*, evidence suggests that culture-specific differences may influence performance. Two cross-cultural studies revealed that exploration strategies in complex problems strongly vary between countries (Güss, Tuason, & Gerhard, 2010; Strohschneider & Güss, 1999). A comparison of university students from Germany, Brazil, India, the Philippines, and the United States showed country-specific differences in performance and exploration behavior in a (computer-based) complex problem solving environment (Güss et al., 2010). Think-aloud techniques and qualitative analyses of verbal protocols were able to show that country-specific problem solving strategies led to differences in the amount of gathered information, the way problems were investigated, and the demonstrated planning and decision making behavior. Reasons for country-specific problem solving styles were seen in "environmental and culture-based differences such as context, resource availability, and emotional expressiveness" (Güss et al., 2010, p. 510). This is in line with the results reported by Wüstenberg et al. (2014)

*Figure 1 (opposite).* Screenshots of the different demands students have to solve within the Genetics Lab, a microworld used to assess complex problem solving. A. Rule Identification: In a fictive genetics lab, students have to manipulate genes (depicted in the diagrams on the left) and identify their effects on a creature's characteristics (depicted in the diagrams on the right). At any time, they can switch to the database to depict the gathered knowledge by clicking on the button in the upper left corner of the screen. B. Rule Knowledge: In a creature related database, students can depict their gathered knowledge by means of a causal diagram. Arrows pointing from genes to characteristics represent a causal effect and indicate the strength (weak or strong) and direction (increasing or decreasing) of this effect. At any time, students can click on the help button in the upper right corner of the screen. C. Rule Application: Students have to apply the gathered knowledge to achieve given target values on the creature's characteristics (indicated by the horizontal lines). Importantly, they only have a limited number of manipulations to do this. Reprinted from "The Genetics Lab: Acceptance and Psychometric Characteristics of a Computer-Based Microworld Assessing Complex Problem Solving," by P. Sonnleitner, M. Brunner, S. Greiff, J. Funke, U. Keller, R. Martin, et al., 2012, *Psychological Test and Assessment Modeling*, 54, p. 59, Figure 1. Copyright 2012 by Pabst Science.



that cross-country differences between Germany and Hungary in the facets of *rule knowledge* and *rule application* were largely attributable to a poor exploration strategy among Hungarian female students, again pointing to the importance of the facet of *rule identification*. Note, however, that no results on the measurement invariance of *rule identification* were reported.

Another reason why performance differences in CPS might be expected for students with immigration background lies in the importance of these skills for the special situation these students are in. According to Martin et al. (2012), problem solving skills are crucial for the academic development of immigrant students, as they might compensate a lack of country-specific prior curricular knowledge or educational experience. In an excellent study drawing on the PISA 2003 data set, Martin et al. supported these claims by showing that nonimmigrant students generally outperformed their immigrant peers on (paper-pencil based) measures of problem solving, mathematics, and science. Moreover, students' problem solving skills were found to strongly relate to students' achievement in mathematics and science, underscoring the cross-curricular importance of problem solving skills for students with immigration background. Similar to academic subjects, the microworlds used to assess CPS are not entirely context-free. Due to the semantic embedding of most microworlds, immigrant students may be disadvantaged if they lack culture-specific knowledge that is (unintentionally) tapped by the administered microworlds. Even an abstract representation of variables or the use of causal diagrams makes strong demands on culture-specific knowledge that might impair immigrant students' performance in these microworlds (see also van de Vijver, 2008, who discussed cross-cultural differences of test understanding and previous test exposure even for reaction-time tests).

A third reason for expected performance differences is based on several studies that have shown that language proficiency in the instructional language predicts performance even in largely "language-independent" subjects, such as mathematics (Kempert, Saalbach, & Hardy, 2011; Levin & Shohamy, 2008). Moreover, when the language in which knowledge acquisition took place differs from the language used for knowledge application, "cognitive costs" arise, impairing students' performance (Kempert et al., 2011; Saalbach, Eckstein, Andri, Hobi, & Grabner, 2013). As microworlds are novel and complex tasks, they usually come with written instructions of varying length (Rollett, 2008). Thus, language proficiency plays an important role in these tasks, even more so when students whose mother tongue differs from the test language are forced to switch language because the language (or "self-talk," i.e., their mother tongue) in which they explore and investigate a problem (see Güss et al., 2010) differs from the language in which the complex problem is presented.

### Possible Benefits of Computer-Based Microworlds for Immigrant Students

In contrast to the factors that might impair immigrant students' performance in CPS tasks, computer-based assessment offers several possibilities that might counter some of these effects or even answer special needs of immigrant students in a way that would not be possible in traditional paper-pencil based assessment instruments. First, microworlds might offer the opportunity to identify immigrant students' cognitive potential despite their educational

background. Conventional (paper-pencil based) measures of higher order cognitive skills, such as reasoning tests, have been found to be positively influenced by attending the academic (i.e., a higher) school track (Becker, Lüdtke, Trautwein, Köller, & Baumert, 2012; Gustafsson, 2008). In such tasks, all the information that is needed to deduce the correct solution is given at the outset and does not need to be generated through interactions with the problem (Rollett, 2008). In contrast, microworlds demand that the students interact with a problem and apply adequate exploration strategies to generate and gather information. Crucially, 81% of Luxembourg students report that they never or hardly ever spend time in laboratories or have to think of appropriate ways to solve science problems by conducting experiments (MENFP, SCRIPT, Université du Luxembourg, & EMACS, 2007). Thus, microworlds might be less affected by educational background than traditional tests, as their task demands are fairly novel and rarely trained in school. This, in turn, might offer the unique opportunity to identify immigrant "underachievers," since immigrant students often attend a nonacademic track that is not appropriate to their cognitive potential due to a lack of language skills (see, e.g., Burton & Martin, 2008, or Klapproth, Glock, Krolak-Schwerdt, Martin, & Böhmer, 2013).

Second, in an attempt to minimize the influence of students' language background on their CPS scores, computer-based assessment makes it possible for each student to take the test in his or her preferred language. In contrast to paper-pencil based measures, a student might even switch the language of the test while working on it. Thus, when CPS is assessed on the computer using a microworld that includes an option to switch to a preferred language, students' language background should not affect performance, or at least affect it to a lesser degree (relative to paper-pencil measures).

Third, the latest generation of microworlds (see, e.g., the Genetics Lab; Sonnleitner et al., 2012; Sonnleitner, Keller, Martin, Latour, & Brunner, in press) avoids extensive written instructions and instead draws on interactive instructions to explain task demands. Since these demands are also illustrated by animations and exercises, instructions only contain a minimum of text. A help-button that is present throughout the students' interaction with the microworld ensures guidance in each phase of the assessment (in students' preferred language). Thus, the impact of language background and proficiency is potentially minimized.

### Aims of the Present Study

Educational systems currently face two important challenges: (a) The requirement to assess cross-curricular competencies such as CPS, which often implies the use of computer-based assessment instruments, and (b) the need to gather empirical knowledge on performance gaps between students with and without immigration background in key cognitive competencies in order to support data-based educational policies. The present article significantly contributes to clarify both issues. First, we tackle the question of whether the Genetics Lab, an established computer-based microworld to assess CPS, is fair with regard to immigration background. This is an essential prerequisite to study any group-related performance differences. Second, drawing on both multidimensional and hierarchical conceptualizations of CPS, we explore performance differences in facets of CPS and general CPS be-



tween students of differing immigration backgrounds. In doing so, the present article is the first to address these key questions on the relations between students' skill to successfully interact with domain-general problems and students' immigration background.

## Method

### Participants

The sample consisted of 299 Luxembourg ninth graders who were enrolled in two different secondary school tracks (i.e., non-academic/intermediate vs. academic school track). Detailed information about the sample is provided in Table 1. One hundred eighty-seven students were enrolled in the nonacademic track (96 of them reporting immigration background), and 112 students were enrolled in the academic track (31 with immigration background). In total, 127 students reported having an immigration background (63 female;  $M$  age = 15.7 years,  $SD$  = 0.81), and 172 students were considered to be native Luxembourg students (83 female;  $M$  age = 15.4 years,  $SD$  = 0.68). Forty-eight of the immigrant students were born abroad with no parent born in Luxembourg (first generation immigrants, 1G;  $n$  = 48), and 79 were born in Luxembourg but reported that both parents were born abroad (second generation immigrants, 2G;  $n$  = 79).

Although previous studies described performance differences between 1G and 2G students (Martin et al., 2012; Stanat, Rauch, & Segeritz, 2010), we pooled these groups for the following reasons. First, similar to 2G students, the vast majority of our sample's 1G students spent their whole academic career in Luxembourg schools. Sixty percent ( $n$  = 27) of 1G students were already enrolled at age 4, when compulsory education in Luxembourg starts, and 77% ( $n$  = 35) were enrolled at age 6, when primary school starts. Thus, with regard to the main characteristic suspected to lead to performance differences, 1G and 2G immigrant students of our sample can be seen as fairly homogenous. Second, a two-samples  $t$  test did not reveal significant differences ( $\alpha$  < 0.05) between 1G and 2G students on the administered cognitive measures (see Table 1 for *means* and *standard deviations*), supporting the notion of homogeneity. Third, the proportion of students that speak one of Luxembourg's official languages at home (i.e., Luxembourgish, French, German) is similar in 1G ( $n$  = 23, 48%) and 2G students ( $n$  = 39, 49%). As language is seen as a crucial factor strongly determining success in the Luxembourg school system (Burton & Martin, 2008; Klapproth et al., 2013), neither of these groups seems to have a related advantage. Note that a comparable number of students in both immigrant groups reported a cultural background similar to Luxembourg, that is they (1G) or both of their parents (2G) were born in neighboring countries (i.e., Belgium, France, Germany). Fourth, by grouping 1G and 2G students, we also ensured comparability with previous Large-Scale Assessment studies that have applied the same grouping (e.g., Schleicher, 2006). Fifth, we gained greater statistical power to detect and support potential effects between immigrant students and their native peers.

### Procedure

The study was conducted with approval from the national Ministry of Education and followed the ethical standards of the host

Table 1  
Means and Standard Deviations for Measures of Reasoning and CPS According to Migration Background and Academic Track

Variable	Without migration background						With migration background					
	Total ( $N$ = 172)			Academic track ( $N$ = 81)			Nonacademic track ( $N$ = 91)			Total ( $N$ = 127)		
	$M$	$SD$	$n$	$M$	$SD$	$n$	$M$	$SD$	$n$	$M$	$SD$	$n$
Cognitive measures												
Reasoning	58.3	17.4		62.2	16.8		54.6	17.3		54.9	16.4	
General CPS	47.4	13.8		50.5	14.6		44.7	12.5		45.4	13.0	
Rule identification	25.6	13.6		28.3	13.5		23.2	13.4		26.9	14.2	
Rule knowledge	64.8	16.4		67.1	16.9		62.7	15.7		61.8	14.8	
Rule application	51.9	18.0		56.2	19.2		48.1	15.9		47.5	16.6	
Sociodemographic characteristics												
Luxembourgish, French or German spoken at home <sup>a</sup>	93	160		95	77		91	83		49	62	
Origin from Belgium, France, or Germany <sup>b</sup>										25	32	
										55	17	
										55	17	
										47	45	
										48	23	
										33	16	
										49	39	
										20	16	

Note. All scores are expressed as POMP scores. CPS = complex problem solving; 1G = first generation; 2G = second generation; POMP scores = percentage of the maximum possible scores.  
<sup>a</sup> Official languages of Luxembourg. <sup>b</sup> Neighboring countries of Luxembourg.

university. A presentation of the study's results and feedback of the students' performance was offered to the volunteering schools. Both students and parents received detailed written information about the scientific background and purpose of the study. None of the students or their parents took the given opportunity to refuse participation.

The Genetics Lab, the reasoning scales, and a background questionnaire were administered by trained research assistants within 110 min (two school lessons) at school during regular class time. To ensure commitment, students were offered a prize for the two best students of each participating class and given detailed written feedback on their performance after completion of the study.

## Measures

**Complex problem solving.** Students' complex problem solving abilities were assessed using the Genetics Lab (see Figure 1), a freely available, computer-based microworld (see <http://www.assessment.lu/GeneticsLab> and Sonnleitner et al., 2012). The Genetics Lab was found to be a reliable and valid measure of CPS that discriminates between and provides reliable scores for the CPS facets *rule identification* (RI), *rule knowledge* (RK), and *rule application* (RA; Sonnleitner et al., 2013).

At the beginning of the Genetics Lab, students could choose between a German, a French, and an English version. Accordingly, instructions and animations were presented in the chosen language. Performance across scenarios was summarized by three scores reflecting students' proficiency in the three main facets of complex problem solving; the scoring algorithms can be found in Keller and Sonnleitner (2012): (a) Each student's exploration strategy was scored on the basis of a detailed *log-file* in which every interaction with the microworld was stored. Thus, it was possible to derive a process-oriented measure (*rule identification*) indicating how efficiently a student explored a scenario by relating the number of informative exploration steps to the total number of steps applied (Kröner et al., 2005). Note that an exploration step is most informative if students manipulate the genes in a way that allows any changes in characteristics to be unambiguously attributed to a certain gene (see Vollmeyer et al., 1996). (b) Students' *rule knowledge* was assessed by scoring their database records (see Figure 1b) using an adaptation of an established scoring algorithm (see Funke, 1992). The resulting *rule knowledge* score thus reflects knowledge about how a gene affects a certain characteristic of a creature and how strong this effect is. (c) Finally, the actions that students took to achieve certain target values on the creature's characteristics during the control phase (see Figure 1c) were used to compute a process-oriented *rule application* score. Only optimal steps (in the sense that the difference from the target values was maximally decreased) were considered to indicate good control performance. Given that all target values must be achieved within three steps, a maximum score of three was possible for each scenario. This approach overcomes the limitations of many previous scoring procedures by guaranteeing that the scoring of a certain control step is completely independent of the preceding control steps. To facilitate the interpretation of the results, all subtest scores were expressed as percentage of the maximum possible score (POMP; see Cohen,

Cohen, Aiken, & West, 1999), for which a value of 0 indicates the lowest possible score, and a value of 100 indicates the highest possible score (see Table 1 for descriptives). All scales showed satisfactory internal consistency, with Cronbach's alpha ranging from .75 for RA to .89 for RI and RK.

When analyzing measurement invariance of facets of CPS, we created parcel scores (i.e., sum scores of subsets of items) in order to better capture the latent constructs. Compared to individual item scores, parcel scores are less prone to distributional violations and show higher reliability (Little, Rhemtulla, & Gibson, in press). Items that shared a theoretical and empirically supported secondary influence beyond the latent construct they were supposed to measure, were combined into a parcel (see Hall, Snell, & Singer Foust, 1999). For each facet, three item parcels were created. Parcel 1 contained items with only two input variables (Items 1, 2, and 3), Parcel 2 contained items with three input variables (Items 4, 5, 6, and 7), and Parcel 3 contained items with three input variables and variables that changed dynamically (Items 8, 9, 10, 11, and 12).

When we conceived of CPS as a hierarchical construct and analyzed measurement invariance of general CPS, we followed the aggregation strategy recommended by Bagozzi and Edwards (1998), using the sum scores of all three subscales as indicators of this general CPS skill.

**Reasoning.** We administered classical paper-pencil measures of reasoning ability to serve as benchmark to explore specific advantages that a computer-based assessment of CPS might have for immigrant students. Specifically, two subtests of the Intelligence Structure Test IST-2000R (Amthauer, Brocke, Liepmann, & Beauducel, 2001), a reliable and valid measure of intelligence, were administered to measure students' ability (a) to complete figural matrix patterns (time limit: 10 min; score MA, Figures 2 and 3), and (b) to complete number series (10 min; score NC, Figures 2 and 3). Both scales were administered in paper-pencil format, and students could choose between a German, French, and an English translation of the instruction. Due to an error in the production process of the test booklets, five out of 20 figural matrix items could not be analyzed. Note that this loss of data was completely at random and affected items of the whole range of difficulty compared to the item difficulties that were obtained for the normative sample (see IST-2000R manual; Amthauer et al., 2001). Analogous to the Genetics Lab, we expressed the score for general reasoning as POMP score with a value of 0 indicating the lowest, and a value of 100 indicating the highest possible score. Whereas internal consistency for number completion was found to be satisfactory ( $\alpha = .90$ ), it was rather poor for the matrices ( $\alpha = .48$ ; compared to  $\alpha = .71$  for the full scale as reported in the manual; Amthauer et al., 2001); this might be explained by the lower number of items. Note, however, that for the analyses of the present study (see below) we did not focus on the internal consistency of manifest scale scores but rather on the factor reliability of the (latent) common reasoning factor underlying both scales, which explained an acceptable 44.3% of their variance (reliability of the reasoning factor score  $\omega = .44$ ; see also Brunner, Nagy, & Wilhelm, 2012).

**Background questionnaire.** A background questionnaire was administered including questions about the students' and their parents' demographic characteristics such as immigration background, language spoken at home, and age.



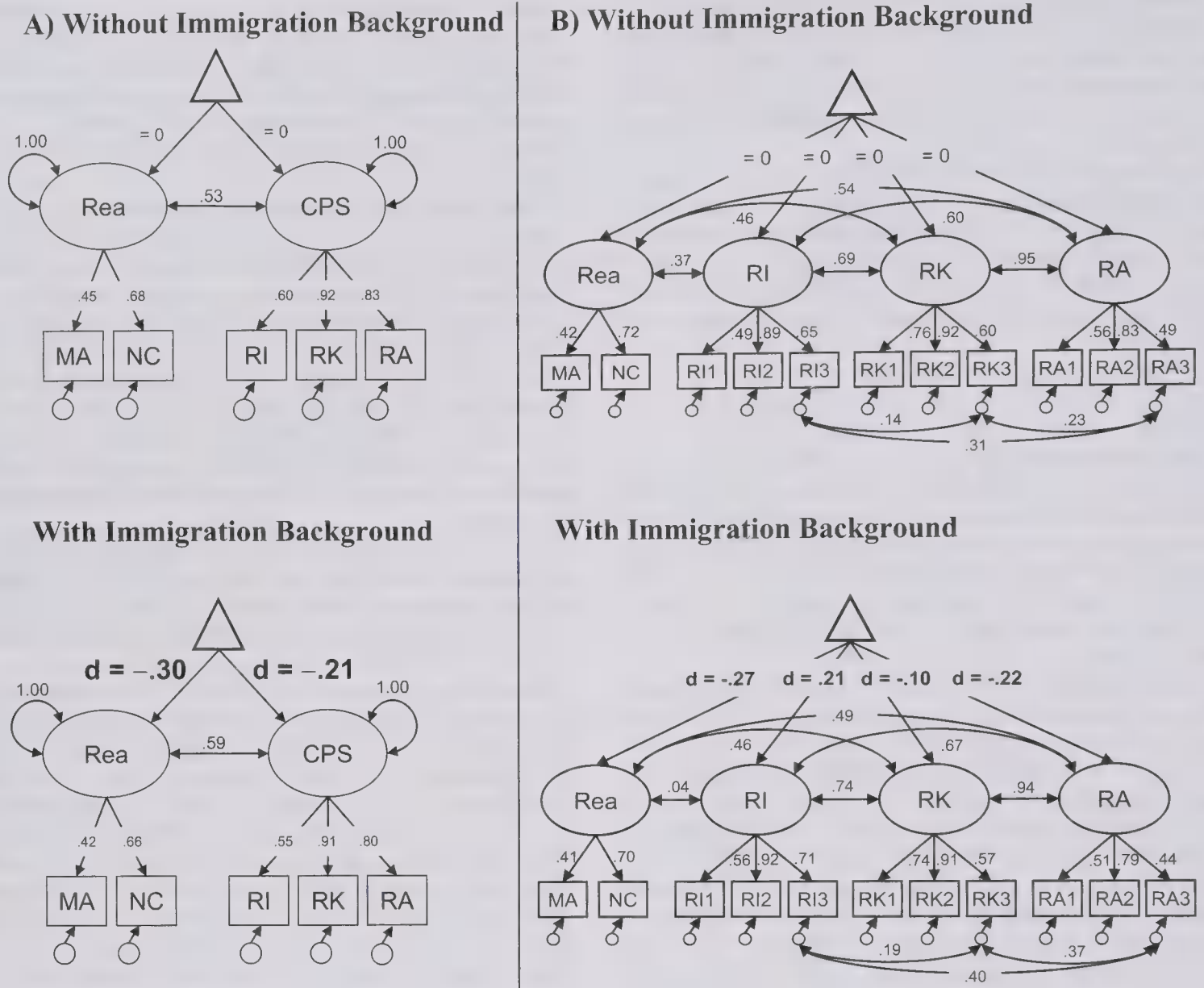


Figure 2. Strict measurement invariant models to study measurement invariance and performance differences in complex problem solving (CPS) and reasoning in relation to students' immigration background. The model in A focuses on general CPS, whereas the model in B considers three different facets of CPS. Rea = reasoning; MA = matrices; NC = number completion; RI = rule identification; RK = rule knowledge; RA = rule application; RI1–RI3 = parcel scores of rule identification items; RK1–RK3 = parcel scores of rule knowledge items; RA1–RA3 = parcel scores of rule application items. Standardized model solution is shown.

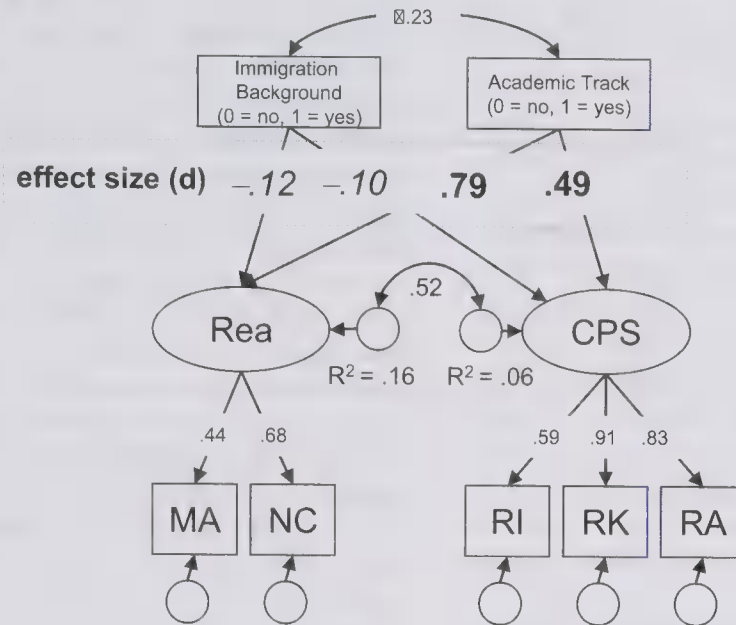
## Statistical Analyses

To study whether students with immigration background differed in their problem solving abilities from their fellow students, we first ensured that the measurement properties of the Genetics Lab and the reasoning tests were invariant across students with differing immigration background. This was done using a stepwise approach based on multiple-group factor analytic models (Little, 1997; Lubke, Dolan, Kelderman, & Mellenbergh, 2003; Widaman & Reise, 1997). Specific levels of measurement invariance (i.e., model constraints that were imposed in the different steps) are explained in the results section. All model parameters were estimated using Mplus 5.2 (L. K. Muthén & Muthén, 1998–2010) by means of the maximum likelihood estimator (ML) and all reported

coefficients are based on standardized solutions. Measurement invariance for a hierarchical conceptualization of CPS including a general CPS skill factor (Figure 2a) was investigated with Models *H1* to *H4*. Measurement invariance for a faceted conceptualization of CPS including the facets of *rule identification*, *rule knowledge*, and *rule application* (Figure 2b) was investigated with Models *F1* to *F4*. The Type I risk  $\alpha$  for data analyses was set at  $p < .05$ , two-tailed.

On the basis of multiple criteria (Little, 1997; Widaman & Reise, 1997), we evaluated whether a certain level of measurement invariance could be assumed for the reasoning scales and the Genetics Lab. First, we consulted the  $\chi^2$  goodness-of-fit statistic and several indices describing overall model fit: the comparative

## A) H-MIMIC



## B) F-MIMIC

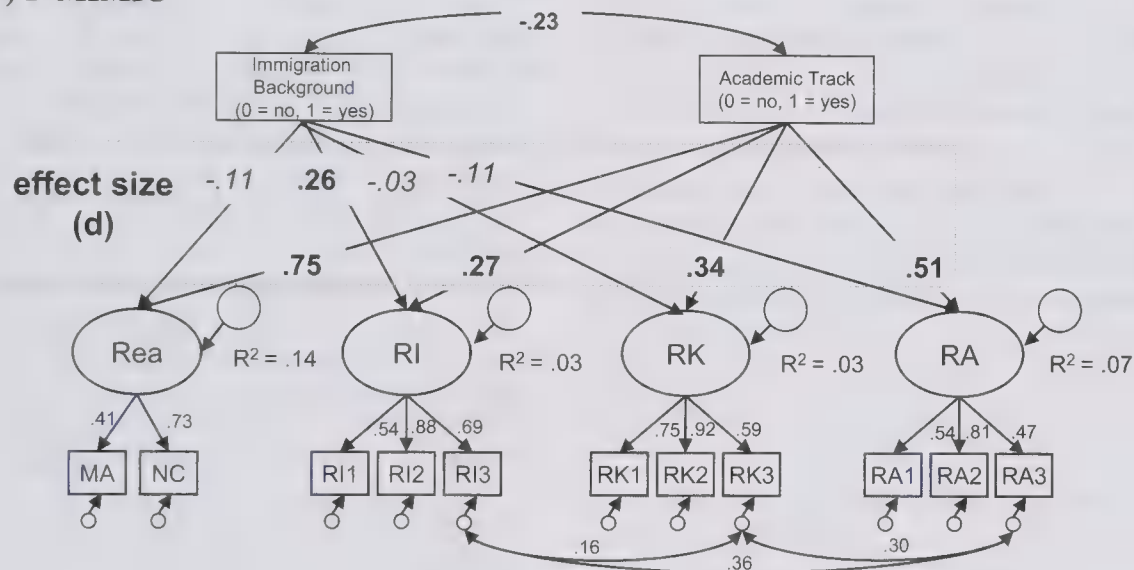


Figure 3. Adjustment of mean differences due to migration background and academic track by means of a multiple-indicator, multiple-causes (MIMIC) model for reasoning and (A) general complex problem solving (CPS) and (B) facets of CPS. For reasons of clarity, not all correlations are shown. H = hierarchical; F = faceted; Rea = Reasoning; MA = matrices; NC = number completion; RI = rule identification; RK = rule knowledge; RA = rule application; RI1–RI3 = parcel scores of rule identification items; RK1–RK3 = parcel scores of rule knowledge items; RA1–RA3 = parcel scores of rule application items. Significant effects are depicted in bold. Standardized model solution is shown.

fit index (CFI); the standardized root-mean-square residual (SRMR); and gamma, which is less sensitive to model size than the related and more popular root-mean-square error of approximation (Fan & Sivo, 2007). For a detailed description of these fit statistics, including formulas, please refer to Iacobucci (2010). Following recommendations by Hu and Bentler (1999), CFI and gamma values above .95 and SRMR values below .08 were considered as indicating good fit between hypothesized model and observed data. In a second step, we checked each model for local misspecifications. Residual correlations above .10 were considered to be problematic (cf. McDonald, 2010). Third, we evaluated whether a more restrictive form of measurement invariance (including addi-

tional cross-group equality constraints) was tenable by inspecting the change in descriptive fit statistics. For CFI, a change of less than .01 was seen as acceptable (Cheung & Rensvold, 2002). For SRMR and gamma, we considered differences below .05 to indicate that cross-group equality constraints had little influence on model fit (see Little, 1997). Since the more restricted measurement models were nested within the less restricted ones, it was further possible to compute  $\chi^2$  difference tests showing whether the models significantly differed in model fit. Fourth, in evaluating the degree of measurement invariance, we also took into account the theoretical implications and parsimony of the models. After consulting the modifications indices of Mplus, we opted for the most



parsimonious and substantive model when local misfit (i.e., the difference between model implied means or covariances on the one side and actual data on the other) was negligible (Little, 1997; McDonald, 2010; Widaman & Reise, 1997).

To determine (a) whether immigration background explained performance discrepancies in students' problem solving abilities and (b) whether such a performance discrepancy was found in both reasoning and CPS, we compared the latent means of students with differing immigration background. For this purpose, we fixed latent means of students without immigration background to zero for identification purposes. Thus, the resulting latent means for students with immigration background represent differences in the construct due to immigration background. Using the obtained latent mean of these students and the pooled standard deviations, which were calculated from the latent standard deviations of both groups, we computed Cohen's  $d$  with positive  $d$  values indicating that students with immigration background outperformed their peers without immigration background.

To this end and to further study the importance of the attended academic track (see above), we ran two multiple-indicator, multiple-causes (MIMIC) models (Jöreskog & Goldberger, 1975; B. O. Muthén, 1989). We adjusted mean differences due to immigration background and school track in a hierarchical (Figure 3, Model H-MIMIC) and a faceted CPS conceptualization (Figure 3, Model F-MIMIC), since the appropriate psychometric conceptualization of CPS is still debated. In both models the unique relations of academic track and immigration background were expressed as standardized effect sizes. Their joint relation to cognitive outcomes was expressed in terms of the amount of explained variance  $R^2$ .

## Results

### Descriptives

Descriptive statistics of the obtained performance scores according to immigration background and academic track are presented in Table 1.<sup>2</sup> As all scores were expressed as percentage of the maximum possible score (POMP; Cohen et al., 1999), a direct comparison between scores and groups is possible when direct inferences on psychometric properties of the scales are avoided. In general, the highest scores were obtained on *rule knowledge*, whereas the lowest scores were obtained on *rule identification*. No performance differences could be found between immigrant students and their nonimmigrant peers in the academic track, but nonimmigrant students slightly outperformed students with immigration background in the nonacademic track. A small difference in favor of immigrant students could be obtained only on the *rule identification* score, indicating that these students applied 3.2% more informative steps when exploring a problem. However, from a practical point of view, these differences are negligible. When comparing performance across school tracks, however, students enrolled in the academic track consistently outperformed students in the nonacademic track. Thus, immigrant students enrolled in the nonacademic track showed the lowest performance on all scales except *rule identification* (see Table 1).

### Measurement Invariance

We first investigated measurement invariance for reasoning and CPS conceived as a hierarchical construct with a general CPS factor (Figure 2a). Model fit indices are given in Table 2. Although the  $\chi^2$  test statistic was found to be significant for the baseline model that assumes the same factorial pattern across groups (i.e., configural measurement invariance; *Model H1*), descriptive fit indices suggested acceptable model fit. Further, no substantive and theoretically justified changes were indicated by local misfit or the modification indices provided by Mplus. Indeed, the observed correlational pattern between reasoning and general CPS mirrored patterns found in previous studies. Thus, configural measurement invariance was tenable for both constructs, and *Model H1* served as benchmark for the subsequent analyses. *Model H2*, assuming weak invariance fitted the data equally well. Descriptive fit indices indicated good fit, and the  $\chi^2$  difference to *Model H1*  $\Delta\chi^2$  was nonsignificant ( $p = .80$ ). Even more restrictive constraints on model parameters in *Model H3* representing strong measurement invariance had little influence on model fit. Compared to *Model H2*, only slight changes were found, with  $\Delta CFI = -.01$  and  $\Delta\chi^2$  remaining nonsignificant. Although strong measurement invariance already ensures a fair comparison of latent factor means and variances across groups (Widaman & Reise, 1997), we further investigated whether strict measurement invariance was tenable (*Model H4*). Model fit indices suggested good fit to the data, change of  $\chi^2$  compared to Model H3 was nonsignificant ( $p = .96$ ) and all parameters were clearly interpretable and of substantive meaning (Figure 2a). Moreover, *Model H4* is the most parsimonious model and modification indices of Mplus were negligible and indicated no local misfit. Thus, we concluded that the most restrictive form of measurement invariance was tenable for reasoning and general CPS, hence allowing for the comparison of latent means on these constructs (but also manifest test scores) for students with differing immigration background.

We also examined measurement invariance when CPS was conceptualized as a faceted construct (Figure 2b). Descriptive fit indices indicated acceptable model fit for our baseline model *F1* representing configural invariance (see Table 2). The same held true for *Model F2*, suggesting that a weak invariant model of a faceted CPS construct was tenable for students with differing immigration background. Note that although  $\chi^2$  was found to be significant for both *Models F1* and *F2*, we emphasized descriptive fit indices since local misfit was negligible. Latent correlations between reasoning and facets of CPS were found to be of the same size as those reported in the literature (Greiff et al., 2013; Kröner et al., 2005; Sonnleitner et al., 2013; Wüstenberg et al., 2012). Moreover, constraining the factor loadings to be equal for students with and without immigration background in *Model F2* did not significantly impact model fit ( $\Delta\chi^2 = 10$  with  $p = .19$  and  $\Delta SRMR = .01$ ). When intercepts were restricted across groups in *Model F3*,  $\Delta\chi^2$  indicated a significant change in model fit. Inspecting descriptive fit statistics, however, revealed only a slight change, with  $\Delta CFI = -.01$  and  $\Delta\gamma = -.01$ . Again we could not identify specific local misfit, and modification indices did not

<sup>2</sup> Note that descriptive statistics for all manifest measures including covariance matrices for immigrant as well as nonimmigrant students are provided in the online supplemental materials.

Table 2

*Measurement Invariance of Reasoning and Complex Problem Solving Conceived Either as Hierarchical or as Faceted Construct*

Model	$\chi^2$	df	p	CFI	SRMR	gamma	$\Delta\chi^2$	$\Delta df$	p	$\Delta CFI$	$\Delta SRMR$	$\Delta gamma$
Measurement invariance of reasoning and hierarchical complex problem solving												
H1. Configural invariance	19	8	.02	.98	.04	.97						
H2. Weak invariance	20	11	.05	.98	.05	.97						
H1 vs. H2							1	3	.80	.00	.01	.00
H3. Strong invariance	28	14	.01	.97	.05	.97						
H2 vs. H3							8	3	.05	-.01	.00	.00
H4. Strict invariance	29	19	.07	.98	.05	.97						
H3 vs. H4							1	5	.96	.01	.00	.00
Measurement invariance of reasoning and faceted complex problem solving												
F1. Configural invariance	123	70	<.01	.96	.06	.94						
F2. Weak invariance	133	77	<.01	.95	.07	.94						
F1 vs. F2							10	7	.19	-.01	.01	.00
F3. Strong invariance	150	84	<.01	.94	.07	.93						
F2 vs. F3							17	7	.02	-.01	.00	-.01
F4. Strict invariance	172	95	<.01	.94	.08	.92						
F3 vs. F4							12	11	.36	.00	.01	-.01
Adjustment of mean differences due to migration background and academic track												
H-MIMIC	19	10	.04	.98	.03	.99						
F-MIMIC	113	49	<.01	.95	.05	.96						

Note. H = hierarchical; F = faceted;  $\chi^2$  = chi-square goodness-of-fit statistic; CFI = comparative fit index; SRMR = standardized root-mean-square residual; MIMIC = multiple-indicator, multiple-causes.

suggest freeing any other parameters. Thus, strong measurement invariance was tenable. Introducing additional across-group constraints on the unique factor invariances or measurement residuals in *Model F4* did not significantly impact model fit ( $\Delta\chi^2 = 12$  with  $p = .36$ ). Moreover, changes in descriptive fit statistics did not exceed .01. As *Model F4* was the most parsimonious of the posited models, and all parameters were substantive (Figure 2b), we concluded that even strict measurement invariance was tenable for CPS as a faceted construct.

Taken together, the analyses of measurement invariance showed that meaningful comparisons for students with differing immigration background could be made for reasoning and CPS scores, regardless of whether the latter were represented as facets of CPS or as a general CPS factor.

### Group Differences Due to Immigration Background

Results given in the lower part of Figure 2a clearly show that students without immigration background significantly outperformed their peers with an immigration background in reasoning ( $d = -0.30$ ) as well as general CPS ( $d = -0.21$ ). However, when group differences were inspected within a faceted conceptualization of CPS (lower part of Figure 2b), the difference did not hold for all facets of CPS equally. Specifically, results obtained for *rule identification* suggested that students with immigration background outperformed their peers when identifying rules and generating knowledge (RI,  $d = 0.21$ ). Despite the effect's small size, this finding is even more remarkable, since the other facets (as could be expected from findings concerning general CPS) are negatively impacted by immigration background, with  $d = -0.10$  for *rule knowledge* and  $d = -0.22$  for *rule application*. In other words, these results suggest that with each subsequent phase of the problem solving process following the generation of knowledge,

performance differences become more pronounced, resulting in the largest performance advantage for native students for *rule application*. Another interesting finding concerns the differential relation of immigration background with reasoning and CPS. Irrespective whether CPS was conceived of as a hierarchical or as faceted construct, measures of CPS seem to be somewhat less affected by immigration background than are measures of reasoning.

### Group Differences Due to Immigration Background and Academic Track

To adjust performance differences with respect to immigration background for differential attendance of the academic track, we ran two MIMIC models that conceptualized CPS either as a hierarchical or faceted construct. Both MIMIC-models showed acceptable model fit (see Table 2). Despite significant  $\chi^2$  statistics, descriptive model fit indices suggested good fit to the data. As we had no indication of local misfits, interpretation of the obtained parameters seemed justified. In general (when controlling for immigration background), we observed that students attending the academic track outperformed students attending the nonacademic track on general CPS, facets of CPS, and reasoning ability (see Figure 3). Crucially, performance differences between students with and without immigration background became negligibly small and nonsignificant when we took the fact into account that students with immigration background were more likely to be enrolled in the nonacademic track (Figure 3). Only when CPS was conceptualized as a faceted construct did a substantial (and statistical significant) influence of immigration background remain for *rule identification* ( $d = 0.26$ ), indicating that students with immigration background outperformed their native peers on this facet of CPS in both academic tracks. As a side note, academic track



explained a considerably larger portion of variance in reasoning ability than in general CPS or facets of CPS, respectively.

### Discussion

Educators across the globe increasingly emphasize the importance of cross-curricular skills (Elliot Bennett et al., 2003; Kuhn, 2009; Ridgway & McCusker, 2003). The importance of these skills has been illustrated through their introduction in several large-scale studies such as PISA (Leutner et al., 2012; OECD, 2010). This suggests that the training of problem solving skills might eventually become an official part of school curricula and that computer-based assessment of complex problem solving will play a central role in the future of educational systems. However, as outlined in the introduction, there are several reasons why one might reasonably expect students' cultural background to influence performance in CPS tasks. The present study is the first to empirically examine (a) whether computer-based assessment of CPS is fair with regard to immigration background, and (b) whether CPS performance differences exist between students with and without immigration backgrounds. To answer these questions, the study drew on a Luxembourg sample of ninth grade students with and without immigration background who were enrolled in different academic tracks.

### Fairness of CPS With Regard to Immigration Background

Several factors can be identified that might affect immigrant students' performance in measures of complex problem solving. Besides culture-specific exploration and knowledge generation strategies (Güss et al., 2010; Strohschneider & Güss, 1999), immigrant students might lack cultural knowledge about the context in which a problem is presented (Martin et al., 2012), and their often poorer language proficiency could impair performance even in largely "language-independent" subjects and assessments (Kempert et al., 2011; Levin & Shohamy, 2008).

However, results of this study clearly showed that the administered microworld to assess CPS (i.e., the Genetics Lab; Sonnleitner et al., 2012) was measurement invariant and thus fair with regard to students' immigration background. As the structure of CPS is still a matter of debate (Sonnleitner et al., 2013), we investigated whether a hierarchical conceptualization including a general CPS factor or a faceted conceptualization including the facets of *rule identification*, *rule knowledge*, and *rule application* could be measured fairly in both groups. Crucially, the highest level of measurement invariance (i.e., strict measurement invariance) could be established for both conceptualizations of CPS (*Models H4* and *F4*). Thus, regardless of whether research is interested in a general CPS skill (e.g., PISA 2013; OECD, 2010) or the facets of CPS (e.g., in applied educational contexts), the obtained performance scores of the Genetics Lab seem to be measurement invariant and thus fair measures of CPS irrespective of students' immigration background. In the light of findings on cross-cultural differences in exploration and knowledge acquisition strategies (Güss et al., 2010; Strohschneider & Güss, 1999), it is even more remarkable that measurement invariance was also tenable for *rule identification*. Given the increase in the numbers of immigrant students in many countries worldwide (Schleicher, 2006), this is an important

prerequisite for comparative small and large-scale studies. This finding also substantially contributes to promising previous results that established measurement invariance for the facets *rule knowledge* and *rule application* with regard to students' sex, nationality (Germany vs. Hungary), and educational background (Greiff et al., 2013; Wüstenberg et al., 2014). The present study is the first, however, showing that measurement invariance is also tenable for the facet *rule identification*, as this facet was not included in previous studies. Nevertheless, given specific features of the Genetics Lab such as game-like characteristics, and multilingual-friendly features (e.g., multilingual, multimedia instructions and a help function), our findings cannot automatically be generalized to other microworlds assessing CPS (see also Greiff et al., 2014).

### Performance Differences With Regard to Immigration Background

In line with previous findings (OECD, 2012; Schleicher, 2006), students with immigration background were generally outperformed by their native peers. Results of *Model H4* (Figure 2a) indicated that native students showed a significantly higher general skill to solve complex problems than their immigrant peers. A comparison with a classic, paper-pencil-based measure of reasoning, however, revealed that immigration background showed a stronger influence on reasoning than on general CPS (Figure 2a).

Crucially, the investigation of a faceted CPS conceptualization (*Model F4*, Figure 2b) was shown to be of substantial value as a possible explanation for the results in *Model H4* concerning general CPS. In the faceted *Model F4*, we could show that immigrant students applied a somewhat more efficient exploration strategy than their nonimmigrant peers. In the subsequent problem solving steps, however, students without immigration background outperformed their immigrant peers. Given these results, it seems that students with immigration background might have difficulties transferring the generated information about a problem into declarative knowledge depicted in causal diagrams or applied to achieve certain target values. Interestingly, the performance gap increased with each phase following the generation of knowledge, resulting in the largest performance difference in the third and final problem solving phase of *rule application*.

Although reasoning tasks as well as CPS tasks draw on cognitive processes that are responsible for the acquisition and the application of knowledge (Wüstenberg et al., 2012), reasoning tasks only provide the final results of these processes. Thus, compared to CPS tasks they might underestimate immigrant students' overall ability. For instance, if *rule application* is most important in determining the correctness of a multiple-choice answer, immigrant students' weakness in transferring and applying the gathered knowledge might result in a stronger effect of immigration background on reasoning scales and hide the strength of immigrant students in *rule identification*. Importantly, this finding clearly shows the value of a faceted conceptualization of CPS and demonstrates that future studies should not only investigate knowledge acquisition or knowledge application or a general CPS skill (as in PISA 2013) but also students' problem exploration strategies.

**Performance differences when taking academic track into account.** On the basis of previous findings showing a strong influence of educational background on CPS skill (e.g., Greiff et



al., 2013) and the fact that the majority of immigrant students in Luxembourg is enrolled in nonacademic tracks (Burton & Martin, 2008; Klapproth et al., 2013), we investigated performance differences in students of differing immigration backgrounds while at the same time controlling for the academic track in which they were enrolled (Figure 3, *Models H-MIMIC* and *F-MIMIC*). Importantly, results of both MIMIC-Models clearly showed that academic track explained most of the differences found for immigration background, with students enrolled in the academic track clearly outperforming their peers in the nonacademic track. Thus, performance differences in CPS and reasoning due to immigration background as found in *Models H4* and *F4* may have simply been due to the fact that the majority of immigrant students were enrolled in the nonacademic track. This finding also indicates that being educated in the academic track may improve performance in these tasks, even if it could also be assumed that initial performance differences might have contributed to the placement decision for the nonacademic track. Note that a positive influence of academic track attendance has also been found for performance on reasoning tasks (Becker et al., 2012; Gustafsson, 2008).

The faceted conceptualization of CPS in *Model F-MIMIC*, however, again provided a substantial finding. Students with immigration background applied more efficient exploration strategies than their native peers irrespective of the academic track they attended. Although descriptive statistics indicated that this difference was only small in size and could only be found in the nonacademic track, manifest measures were not free from measurement error and may thus underestimate real (latent) differences. For individual assessment, this finding is of special importance, as it may point to otherwise overlooked potential. As mentioned above, especially in Luxembourg, many students with immigration background are oriented to nonacademic tracks due to low language proficiency in any or all of the country's three official languages, Luxemburgish, French, and German (Burton & Martin, 2008; Klapproth et al., 2013; Shewbridge, Tamassia, Santiago, & Ehren, 2012). Given the positive influence of academic track on reasoning scales found in our study and reported by Becker et al. (2012) and Gustafsson (2008), such scales may not be suited to assess students' real cognitive potential; they might simply reflect the training that is (or is not) provided in a specific academic track. Importantly, we still found a significant relation between immigration background and *rule identification* after controlling for academic track, indicating that educational background had less influence on this skill. Since systematic exploration behavior is presumably trained less extensively in school than thinking in causal diagrams or the application of knowledge, novelty of task demands might explain this finding. This is possibly reflected in the generally lower performance-scores in *rule identification* compared to the other assessed abilities (see Table 1). Consequently, *rule identification* might be used to identify immigrant "underachievers" who are enrolled in non-academic tracks due to their lower language proficiency but have the cognitive potential to succeed in an academic track. In sum, results suggest that CPS may be a fairer and more valid assessment of students' cognitive abilities than traditional reasoning scales, and this may in particular hold for the dimension of *rule identification*.

## Limitations and Outlook

A crucial aspect in investigating immigrant students is the definition of immigration background. For the present article, we applied an internationally well-established classification, defining students as having immigrant status if they were either born abroad and later moved to the host country or they were born in the host country but both parents were born abroad (cf. OECD, 2012). Despite broad acceptance of this term, there is still ambiguity concerning students who have only one parent born abroad or parents who were born in two different countries. In these cases, a classification concerning immigrant status or country of origin is difficult (Stanat et al., 2010). This, however, highlights the heterogeneity of most immigrant student samples. Nevertheless, from an educational or even political perspective, focus should be set on analyzing groups that are homogeneous with regard to performance, as they allow broader conclusions for everyday educational practice. The grouping of first- and second-generation immigrant students in our study led to such a (relatively) homogeneous sample with regard to cognitive performance.

The current investigation of measurement invariance and students' performance differences with regard to immigration background would undoubtedly have profited from a larger sample. Although the percentage of immigrant students in both school tracks can be seen as representative for Luxembourg (Burton & Martin, 2008), size of subsamples was too low to conduct more fine-grained analyses, for example concerning the effect of country of origin. Note, however, that sample size was large enough to detect statistically significant effects for the group as a whole. Nevertheless, this study should only be seen as a starting point. Further research is needed to generalize our findings to different school grades and countries or to study specific subsamples of immigrant students with special risk factors. Such studies would undoubtedly be beneficial for educational policies aimed at reducing the performance gap between immigrant students and their native peers. The upcoming PISA 2012 data set (OECD, 2010), including measures on CPS and detailed information on immigration background, might be a rich source for such analyses.

Another limitation concerns the interpretation of the observed effects of immigrant background and school track. Since random allocation is impossible in such contexts, it is not possible to establish causality. Positive effects of academic track can either be interpreted as positive influence of the advanced training on problem solving skills or simply as evidence that students with better problem solving skills are more likely to attend an academic track. Longitudinal studies would provide interesting information on such schooling effects but would nevertheless be limited by non-random allocation of students.

Although a slight advantage in *rule identification* could be identified for immigrant students, it is still unclear why these differences occur. More fine-grained analyses of the exploration strategies shown in students' log-file data (Hadwin, Nesbit, Jamieson-Noel, Code, & Winne, 2007) or even think-aloud protocols might shed light on this issue, as such studies would allow understanding the exact strategies that students use in order to identify the underlying rules. Previous qualitative studies have shown, for example, that German university students employ more control-oriented strategies in CPS tasks than students from India or Brazil (Güss et al., 2010; Strohschneider & Güss, 1999). A similar



effect might explain our observed differences in *rule identification*, though the results of existing studies cannot automatically be generalized to our study, since the problem solving strategies applied by students depend strongly on the specific microworld used to assess CPS.

## Conclusion

In sum, the present study provides evidence that CPS as assessed by the Genetics Lab, a computer-based microworld that incorporates multilingual-friendly features (e.g., multilingual, multimedia instructions, and a help function) can be equitably measured with respect to students' immigration background. Such fairness is a prerequisite for small and large-scale studies (such as PISA) that aim to compare complex problem-solving abilities of students with differing immigration backgrounds. Special value was shown in the analysis of students' exploration strategies, highlighting the informative potential of computer-derived process measures and pointing to a future direction in CPS research. This last point highlights the value of a faceted approach for CPS, which might have the potential to reveal especially in immigrant students cognitive facets that can be considered to be a relative strength for these students and that might go unnoticed by educators without the existence of adequate CPS assessments. Thus, CPS assessment shows a high potential for being a central element in future educational curricula with a strong focus on cross-curricular skills.

## References

- Abele, S., Greiff, S., Gschwendtner, T., Wüstenberg, S., Nickolaus, R., Nitzschke, A., & Funke, J. (2012). Dynamische Problemlösekompetenz [Dynamic problem solving]. *Zeitschrift für Erziehungswissenschaft*, 15, 363–391. doi:10.1007/s11618-012-0277-9
- Anthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *Intelligenz-Struktur-Test 2000 R* [Intelligence Structure Test 2000 R]. Göttingen, Germany: Hogrefe.
- Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods*, 1, 45–87. doi:10.1177/109442819800100104
- Becker, M., Lüdtke, O., Trautwein, U., Köller, O., & Baumert, J. (2012). The differential effects of school tracking on psychometric intelligence: Do academic-track schools make students smarter? *Journal of Educational Psychology*, 104, 682–699. doi:10.1037/a0027608
- Blech, C., & Funke, J. (2005). *Dynamis review: An overview about applications of the Dynamis approach in cognitive psychology*. Retrieved from [http://www.die-bonn.de/espid/dokumente/doc-2005/blech05\\_01.pdf](http://www.die-bonn.de/espid/dokumente/doc-2005/blech05_01.pdf)
- Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, 80, 796–846. doi:10.1111/j.1467-6494.2011.00749.x
- Buchner, A. (1995). Basic topics and approaches to the study of complex problem solving. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 27–63). Hillsdale, NJ: Erlbaum.
- Bühner, M., Kröner, S., & Ziegler, M. (2008). Working memory, visual-spatial-intelligence and their relationship to problem-solving. *Intelligence*, 36, 672–680. doi:10.1016/j.intell.2008.03.008
- Burton, R., & Martin, R. (2008). L'orientation scolaire au Luxembourg: "Au-delà de l'égalité des chances . . . le gâchis d'un potentiel humain" [Tracking decisions in Luxembourg: "Beyond equal opportunities . . . a waste of human capital"]. In R. Martin, C. Dierendonck, C. Meyers, & M. Noesen (Eds.), *La place de l'école dans la société luxembourgeoise de demain* (pp. 165–186). Bruxelles, Belgium: DeBoeck.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255. doi:10.1207/S15328007SEM0902\_5
- Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research*, 34, 315–346. doi:10.1207/S15327906MBR3403\_2
- Elliot Bennett, R., Jenkins, F., Persky, H., & Weiss, A. (2003). Assessing complex problem solving performances. *Assessment in Education: Principles, Policy & Practice*, 10, 347–359.
- Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, 42, 509–529. doi:10.1080/00273170701382864
- Funke, J. (1992). Dealing with dynamic systems: Research strategy, diagnostic approach and experimental results. *German Journal of Psychology*, 16, 24–43.
- Funke, J. (2003). *Problemlösendes Denken* [Problem solving thinking]. Stuttgart, Germany: Kohlhammer.
- Funke, J. (2010). Complex problem solving: A case for complex cognition? *Cognitive Processing*, 11, 133–142. doi:10.1007/s10339-009-0345-0
- Greiff, S., Kretzschmar, A., Müller, J., Spinath, B., & Martin, R. (2014). Computer-based assessment of complex problem solving in educational contexts and how it is influenced by students' level of information and communication technology literacy. *Journal of Educational Psychology*, 106, 666–680. doi:10.1037/a0035426
- Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement*, 36, 189–213. doi:10.1177/0146621612439620
- Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts—something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology*, 105, 364–379. doi:10.1037/a0031856
- Güss, C. D., Tuason, M. T., & Gerhard, C. (2010). Cross-national comparisons of complex problem-solving strategies in two microworlds. *Cognitive Science*, 34, 489–520. doi:10.1111/j.1551-6709.2009.01087.x
- Gustafsson, J.-E. (2008). Schooling and intelligence: Effects of track of study on level and profile of cognitive abilities. In P. Kyllonen, R. Roberts, & L. Stankov (Eds.), *Extending intelligence: Enhancement and new constructs* (pp. 37–59). Mahwah, NJ: Erlbaum.
- Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., & Winne, P. H. (2007). Examining trace data to explore self-regulated learning. *Metacognition and Learning*, 2, 107–124. doi:10.1007/s11409-007-9016-7
- Hall, R. J., Snell, A. F., & Singer Foust, M. (1999). Item parceling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. *Organizational Research Methods*, 2, 233–256. doi:10.1177/109442819923002
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55. doi:10.1080/10705519909540118
- Iacobucci, D. (2010). Structural equations modeling: Fit Indices, sample size, and advanced topics. *Journal of Consumer Psychology*, 20, 90–98. doi:10.1016/j.jcps.2009.09.003
- Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 10, 631–639.
- Keller, U., & Sonnleitner, P. (2012). *Genetics Lab scoring algorithm*. Luxembourg, Luxembourg: University of Luxembourg.
- Kempert, S., Saalbach, H., & Hardy, I. (2011). Cognitive benefits and costs of bilingualism in elementary school students: The case of mathematical word problems. *Journal of Educational Psychology*, 103, 547–561. doi:10.1037/a0023619
- Klapproth, F., Glock, S., Krolak-Schwerdt, S., Martin, R., & Böhmer, M. (2013). Prädiktoren der Sekundarschulempfehlung in Luxemburg [Predictors of recommendations for secondary school type in Luxembourg].

- Zeitschrift für Erziehungswissenschaft*, 16, 355–379. doi:10.1007/s11618-013-0340-1
- Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence*, 33, 347–368. doi:10.1016/j.intell.2005.03.002
- Kuhn, D. (2009). Do students need to be taught how to reason? *Educational Research Review*, 4, 1–6. doi:10.1016/j.edurev.2008.11.001
- Leutner, D., Fleischer, J., Wirth, J., Greiff, S., & Funke, J. (2012). Analytische und dynamische Problemlösekompetenz im Lichte internationaler Schulleistungsvergleichsstudien [Analytic and dynamic problem solving competence in the light of international student large scale assessment]. *Psychologische Rundschau*, 63, 34–42. doi:10.1026/0033-3042/a000108
- Levin, T., & Shohamy, E. (2008). Achievement of immigrant students in mathematics and academic Hebrew in Israeli school: A large-scale evaluation study. *Studies in Educational Evaluation*, 34, 1–14. doi:10.1016/j.stueduc.2008.01.001
- Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, 32, 53–76. doi:10.1207/s15327906mbr3201\_3
- Little, T. D., Rhemtulla, M., & Gibson, K. (in press). Why the items versus parcels controversy needn't be one. *Psychological Methods*.
- Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence*, 31, 543–566. doi:10.1016/S0160-2896(03)00051-5
- Martin, A. J., Liem, G. A. D., Mok, M. M. C., & Xu, J. (2012). Problem solving and immigrant student mathematics and science achievement: Multination findings from the Programme for International Student Assessment (PISA). *Journal of Educational Psychology*, 104, 1054–1073. doi:10.1037/a0029152
- McDonald, R. P. (2010). Structural models and the art of approximation. *Perspectives on Psychological Science*, 5, 675–686. doi:10.1177/1745691610388766
- MENFP, SCRIPT, Université du Luxembourg, & EMACS. (2007). *PISA 2006: Rapport national Luxembourg* [PISA 2006: National report Luxembourg]. Luxembourg, Luxembourg: Ministère de l'Éducation nationale et de la Formation professionnelle – Service de Coordination de la Recherche et de l'Innovation Pédagogiques et Technologiques and Université du Luxembourg.
- Meunier, M. (2011). Immigration and student achievement: Evidence from Switzerland. *Economics of Education Review*, 30, 16–38. doi:10.1016/j.econedurev.2010.06.017
- Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, 54, 557–585. doi:10.1007/BF02296397
- Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide* (5th ed). Los Angeles, CA: Muthén & Muthén.
- OECD. (2010). *PISA 2012 problem solving framework (draft for field trial)*. Paris, France: OECD.
- OECD. (2012). *Untapped skills: Realising the potential of immigrant students*. Paris, France: OECD.
- Ridgway, J., & McCusker, S. (2003). Using computers to assess new educational goals. *Assessment in Education: Principles, Policy & Practice*, 10, 309–328.
- Rollett, W. (2008). *Strategieinsatz, erzeugte Information und Informationsnutzung bei der Exploration und Steuerung komplexer dynamischer Systeme* [Strategy use, generated information and use of information in exploring and controlling complex, dynamic systems]. Berlin, Germany: Lit Verlag.
- Saalebach, H., Eckstein, D., Andri, N., Hobi, R., & Grabner, R. H. (2013). When language of instruction and language of application differ: Cognitive costs of bilingual mathematics learning. *Learning and Instruction*, 26, 36–44. doi:10.1016/j.learninstruc.2013.01.002
- Schleicher, A. (2006). Where immigrant students succeed: A comparative review of performance and engagement in PISA 2003. *Intercultural Education*, 17, 507–516. doi:10.1080/14675980601063900
- Schweizer, F., Wüstenberg, S., & Greiff, S. (2013). Validity of the Micro-DYN approach: Complex problem solving predicts school grades beyond working memory capacity. *Learning and Individual Differences*, 24, 42–52. doi:10.1016/j.lindif.2012.12.011
- Shewbridge, C., Tamassia, C., Santiago, P., & Ehren, M. (2012). *OECD reviews of evaluation and assessment in education: Luxembourg 2012*. Paris, France: OECD. doi:10.1787/9789264116801-en
- Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., . . . Latour, T. (2012). The Genetics Lab: Acceptance and psychometric characteristics of a computer-based microworld assessing complex problem solving. *Psychological Test and Assessment Modeling*, 54, 54–72.
- Sonnleitner, P., Keller, U., Martin, R., & Brunner, M. (2013). Students' complex problem-solving abilities: Their structure and relations to reasoning ability and educational success. *Intelligence*, 41, 289–305. doi:10.1016/j.intell.2013.05.002
- Sonnleitner, P., Keller, U., Martin, R., Latour, T., & Brunner, M. (in press). Assessing complex problem solving in the classroom: Meeting challenges and opportunities. In B. Csapó & J. Funke (Eds.), *The nature of problem solving* (pp. xx–xx). Paris, France: OECD.
- Stanat, P., Rauch, D., & Segeritz, M. (2010). Schülerinnen und Schüler mit Migrationshintergrund [Students with immigration background]. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, . . . P. Stanat (Eds.), *PISA 2009. Bilanz nach einem Jahrzehnt* (pp. 200–230). Münster, Germany: Waxmann.
- Strohschneider, S., & Güss, D. (1999). The fate of the Moros: A cross-cultural exploration of strategies in complex and dynamic decision making. *International Journal of Psychology*, 34, 235–252. doi:10.1080/0020759993999873
- van de Vijver, F. J. R. (2008). On the meaning of cross-cultural differences in simple cognitive measures. *Educational Research and Evaluation*, 14, 215–234. doi:10.1080/13803610802048833
- Vollmeyer, R., Burns, B. D., & Holyoak, K. J. (1996). Impact of goal specificity on strategy use and acquisition of problem structure. *Cognitive Science*, 20, 75–100. doi:10.1207/s15516709cog2001\_3
- Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association. doi:10.1037/10222-009
- Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving—More than reasoning? *Intelligence*, 40, 1–14. doi:10.1016/j.intell.2011.11.003
- Wüstenberg, S., Greiff, S., Molnár, G., & Funke, J. (2014). Cross-national gender differences in complex problem solving and their determinants. *Learning and Individual Differences*, 29, 18–29. doi:10.1016/j.lindif.2013.10.006

Received April 30, 2013

Revision received October 18, 2013

Accepted November 17, 2013 ■



## Boredom and Academic Achievement: Testing a Model of Reciprocal Causation

Reinhard Pekrun  
University of Munich

Nathan C. Hall  
McGill University

Thomas Goetz  
University of Konstanz and  
Thurgau University of Teacher Education

Raymond P. Perry  
University of Manitoba

A theoretical model linking boredom and academic achievement is proposed. Based on Pekrun's (2006) control-value theory of achievement emotions, the model posits that boredom and achievement reciprocally influence each other over time. Data from a longitudinal study with college students ( $N = 424$ ) were used to examine the hypothesized effects. The study involved 5 assessments of students' boredom and test performance during a university course spanning an entire academic year. Structural equation modeling was used to examine effects of boredom on achievement, and vice versa. The results show that boredom had consistently negative effects on subsequent performance, and performance had consistently negative effects on subsequent boredom, while controlling for students' gender, age, interest, intrinsic motivation, and prior achievement. These results provide robust evidence for the proposed links between boredom and achievement and support systems-theoretical perspectives on the dynamics of emotions and achievement. From a broader educational perspective, the findings imply that researchers and practitioners alike should focus attention on boredom as an important, yet often overlooked, academic emotion.

**Keywords:** boredom, achievement emotion, academic achievement, control-value theory, intrinsic motivation

Research on achievement emotions has largely focused on test anxiety (Zeidner, 1998, 2007). Emotions other than anxiety were neglected, despite their ubiquity in the classroom and relevance to students' academic performance, development, and health (Pekrun, Goetz, Titz, & Perry, 2002). Recent progress in educational research on emotions has begun to draw attention to the diversity of emotions that students experience (Linnenbrink-Garcia & Pekrun, 2011; Pekrun & Linnenbrink-Garcia, in press; Schutz & Lanehart, 2002; Schutz & Pekrun, 2007). However, surprisingly few of these studies have featured students' boredom.

Outside academia, boredom has been shown to relate to delinquency, gambling, drug use, and health problems (Amos, Wiltshire, Haw, & McNeill, 2006; Blaszczynski, McConaghy, & Frankova, 1990; Newberry & Duncan, 2001; Thackray, 1981). Thus, it seems plausible to assume that boredom can have equally pronounced effects within achievement settings. Boredom is an emotion that is among the most frequently experienced, and potentially most devastating, affective states occurring in the classroom (Mann & Robinson, 2009; Pekrun, Goetz, Daniels, Stupnisky, & Perry, 2010).

Accordingly, the present research examined the relationship between students' boredom and their academic achievement. It is posited that these two constructs are related by reciprocal causation, in contrast to traditional unidirectional models of emotions affecting achievement. Sporadic correlational evidence suggests that boredom is negatively related to academic achievement, as detailed below. However, the underlying causal relationships are virtually unexplored. Hence, the reciprocal linkages between these two constructs have not been examined, leaving unresolved the question whether boredom is functionally relevant in impacting performance, or is just an epiphenomenon of low achievement.

Examining reciprocal relations between boredom and achievement is of considerable theoretical and practical importance. Evidence on reciprocal relations bears on the validity of systems-oriented theories proposing that emotions and performance are linked by feedback loops rather than by unidirectional causation (Pekrun, 2006; Turner & Waugh, 2007). With regard to educa-

---

This article was published Online First February 24, 2014.

Reinhard Pekrun, Department of Psychology, University of Munich, Munich, Germany; Nathan C. Hall, Department of Educational and Counselling Psychology, McGill University, Montreal, Quebec, Canada; Thomas Goetz, Department of Empirical Educational Research, University of Konstanz, Konstanz, Germany; and Thurgau University of Teacher Education, Thurgau, Switzerland; Raymond P. Perry, Department of Psychology, University of Manitoba, Winnipeg, Manitoba, Canada.

This research was supported by a Research Chair Grant awarded to Reinhard Pekrun by the University of Munich, a Postdoctoral Fellowship Award from the Humboldt Foundation to Nathan C. Hall, and a Konrad Adenauer Award from the Royal Society of Canada and the Humboldt Foundation to Raymond P. Perry.

Correspondence concerning this article should be addressed to Reinhard Pekrun, Department of Psychology, University of Munich, Leopoldstrasse 13, 80802 Munich, Germany. E-mail: pekrun@lmu.de

tional practice, it is important for teachers, administrators, and parents to know whether boredom is detrimental to students' performance, or simply a by-product of poor achievement that is of secondary relevance because it does not impact future educational attainment.

In the following sections, we first discuss the construct of boredom as an achievement emotion. We then summarize the paucity of evidence on boredom and achievement and present a theoretical model on their reciprocal linkages. The model was tested in an empirical study that investigated course-related boredom and course performance in undergraduate students over a full academic year.

### Boredom as an Achievement Emotion

Boredom comprises unpleasant feelings, reduced physiological arousal, perceived lack of cognitive stimulation, task-irrelevant thinking (e.g., daydreaming), prolonged subjective duration of time, and impulses to escape the boredom-inducing situation through disengagement (Goetz & Hall, in press; Mikulas & Vondanovich, 1993; Pekrun et al., 2010; van Tilburg & Igou, 2012; Vogel-Walcutt, Fiorella, Carper, & Schutz, 2012; for variants of the boredom experience, see Goetz et al., in press). From an evolutionary perspective, boredom serves to limit engagement in activities that lack consummatory value, that do not promise to yield any reinforcement, and that are not suited to broaden the individual's thought-action repertoire, thus making it possible to redirect attention toward more rewarding activities.

Boredom is experienced frequently by students in academic achievement settings (Larson & Richards, 1991; Mann & Robinson, 2009; Pekrun et al., 2010). In contrast to emotions linked to success and failure outcomes (e.g., anxiety, pride, or shame), boredom relates to the achievement activities performed in these settings, such as attending classes or doing homework. Achievement emotions are defined as emotions related to achievement activities or their outcomes (Pekrun, 2006) and can be classified using the  $2 \times 2$  (Object Focus  $\times$  Valence) taxonomy of achievement emotions proposed by Pekrun et al. (2002). Within this taxonomy, boredom represents an activity-related achievement emotion since the reference object is the current achievement activity, rather than the outcome of the activity.

As boredom involves low arousal and distinct affective, cognitive, and motivational components as described earlier, it differs from other negative emotions such as anger, test anxiety, or shame (van Tilburg & Igou, 2012). Correlations between boredom and other negative emotions such as test anxiety have been in the range of  $r = .30$  to  $.50$ , thus corroborating the distinctiveness of this emotion (Daniels et al., 2009; Pekrun, Goetz, Frenzel, Barchfeld, & Perry, 2011; van Tilburg & Igou, 2012). Moreover, boredom is also different from a mere lack of positive affect, situational interest, or flow experiences (Csikszentmihalyi, 1975). Boredom is an unpleasant affective state that consists of specific component processes that can be highly aversive, thus being more than simply the absence of positive affect.

Regarding situational interest (Hidi & Renninger, 2006), boredom can arise from a lack of interest but is not identical with it (Pekrun et al., 2010). Situational interest, as well as lack of interest, can relate to various positive and negative emotions (e.g., enjoyment, disgust) but are distinct from any specific single emo-

tion (for a review, see Ainley & Hidi, in press). As such, lack of situational interest need not be combined with boredom. For example, when preparing for an exam, a student may lack interest in the material but may feel panic and a fear of failing the exam rather than boredom. From a motivational perspective, lack of interest is conceptually equivalent to a lack of approach motivation, whereas boredom is equivalent to avoidance motivation in wishing to escape the situation, which implies that lack of interest and boredom belong to different categories of affect.

### Previous Research on Boredom and Academic Achievement

In this section, we summarize the available evidence on linkages between students' boredom and their academic achievement. Given the dearth of direct evidence on the boredom-achievement link, we also consider more indirect evidence arising from findings on boredom and students' ability and engagement underlying achievement.

#### Boredom and Academic Achievement

Maroldo (1986) and Pekrun et al. (2010) found that college students' boredom correlated negatively with their grade point average. Similarly, research with middle and high school students also found boredom to correlate negatively with academic achievement. Frenzel, Pekrun, and Goetz (2007) reported that fifth to 10th graders' boredom during math classes related negatively to their math achievement. In an investigation of domain-specific achievement emotions, boredom and academic grades in different school subjects correlated negatively in samples of eighth and 11th graders (Goetz, Frenzel, Pekrun, Hall, & Lüdtke, 2007). An exception to this pattern of uniformly negative correlations is a study reported by Larson and Richards (1991). Using an experience sampling method to assess fifth to ninth graders' boredom during schoolwork, this study found small positive correlations between boredom and students' achievement ( $r_s = .15$  and  $.13$  for grade point average [GPA] and test scores, respectively).

Whereas all these studies were cross-sectional, the studies reported by Ahmed, van der Werf, Kuyper, and Minnaert (2013); Pekrun, Elliot, and Maier (2009); and Pekrun et al. (2010, Study 5) used longitudinal designs. Ahmed et al. (2013) found that seventh graders' boredom was negatively related to their math achievement, and that change in boredom over one school year was negatively related to concurrent change of math achievement. In Pekrun et al.'s (2009) investigation, undergraduates' boredom arising from studying for a course exam was a negative predictor of performance on the exam, and Pekrun et al. (2010) found that students' boredom in a university course negatively predicted their end-of-year course performance. These studies suggest that boredom is a negative predictor of students' academic performance, but they did not examine reverse effects of performance on boredom. In sum, students' boredom experienced in academic achievement settings has almost uniformly been found to correlate negatively with their achievement; however, evidence on reverse effects of achievement on boredom, and on reciprocal relations between the two constructs over time, is lacking.



## Boredom and Cognitive Ability

Traditionally, it was assumed that boredom is caused by a lack of challenge resulting from a combination of high individual ability and low task demands (Csikszentmihalyi, 1975). In the educational literature, boredom was attributed to gifted children dealing with environments tailored to the needs of average-ability students ("The bored and disinterested gifted child"; Sisk, 1988, p. 5; also see Rennert & Berger, 1956). By contrast, the evidence from a few survey studies suggests that boredom is more frequently experienced by low-ability individuals. Roseman (1975) found that bored students were overrepresented among middle school students having IQ scores less than 95. Similarly, Fogelman (1976) showed that 11-year-olds who reported being "often bored" in their spare time had significantly lower cognitive abilities than students who were "sometimes bored" or "always enjoyed" their leisure time. Congruent with lower ability scores, bored students in middle school also reported lower self-concepts of ability (Goetz, Pekrun, Hall, & Haag, 2006). This evidence suggests that both objective and perceived cognitive ability are negatively related to students' boredom.

## Boredom and Achievement Behavior

A few studies indicate that boredom relates negatively to students' attention and effort in achievement activities. Farmer and Sundberg (1986) reported that undergraduates' boredom proneness correlated negatively with their attentiveness during lectures. Similarly, Mann and Robinson (2009) found that university students' boredom related to their off-task behavior during lectures and to missing future lectures. Watt and Vodanovich (1999) demonstrated that college students' boredom related negatively to their educational involvement and career planning. Pekrun et al. (2002, 2010) reported that university students' boredom was negatively related to their attention, effort, self-regulation of learning, and use of flexible learning strategies.

Similarly, boredom relates negatively to academic engagement in middle and high school students. Based on interviews with sixth and seventh grade students, Jarvis and Seifert (2002) found that students withdrew effort at school because of being bored. In Roseman's (1975) investigation, students' boredom related negatively to teacher and parent ratings of how hard students worked. Skinner, Furrer, Marchand, and Kindermann (2008) reported that fourth to seventh graders' boredom correlated negatively with their behavioral engagement. Finally, Ahmed et al. (2013) found that seventh grade students' boredom in mathematics related negatively to their use of learning strategies. Overall, the findings indicate that boredom relates negatively to students' attention, investment of effort, and self-regulation of learning.

## Summary of Previous Research

In sum, boredom has been found to correlate negatively with students' academic achievement as well as related variables including cognitive ability, academic self-concepts, and behavioral engagement. By contrast, the extant research does not support the classical notion that high-achieving students suffer more from boredom in achievement settings, compared with average-achieving or low-achieving students. Rather, the evidence suggests that boredom is linked to low achievement.

However, as most of the available evidence is cross-sectional, causal conclusions are not warranted. Correlations between boredom and achievement leave unanswered the question of whether boredom reduces achievement, low achievement causes boredom, or both are correlated due to third variables. For examining reciprocal effects, panel designs would be needed that measure both boredom and performance at multiple points of time, thus making it possible to analyze effects of one variable on the other while controlling for previous levels of both variables (McArdle, 2009). No analysis of this type is available to date. As yet, there is no empirical answer to address how boredom and achievement reciprocally influence each other, and how these influences unfold over time.

## Theoretical Framework: A Reciprocal Effects Model of Boredom and Achievement

Hypotheses regarding linkages between boredom and achievement were derived from Pekrun's (2006; Pekrun & Perry, 2013) control-value theory of achievement emotions. This theory integrates propositions from expectancy-value, attributional, and control approaches to achievement emotions (Folkman & Lazarus, 1985; Pekrun, 1992; Turner & Schallert, 2001; Weiner, 1985). It expands upon these approaches by addressing not only outcome emotions, but activity emotions such as boredom as well. The theory posits that achievement emotions are aroused by cognitive appraisals of control over, and the subjective value of, achievement activities and their outcomes. Control appraisals consist of perceptions of one's ability to successfully perform actions (i.e., self-efficacy expectations) and to attain outcomes (outcome expectations). Value appraisals pertain to the perceived importance of these activities and outcomes. Furthermore, the theory posits that these emotions in turn influence achievement behavior and performance. Since performance outcomes shape succeeding perceptions of control over performance, one important implication is that emotions, their appraisal antecedents, and their performance outcomes are linked by reciprocal causation. In terms of reciprocal causation, the theory is consistent with reciprocal effects models for variables such as students' self-concepts (Marsh & Craven, 2006; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005), achievement goals (Linnenbrink & Pintrich, 2002), interest (Hacker, Durik, Barron, Linnenbrink-Garcia, & Tauer, 2008), and anxiety (Pekrun, 1992).

## Effects of Boredom on Achievement

The control-value theory proposes that achievement emotions influence students' cognitive resources, motivation to learn, and use of learning strategies. Boredom is expected to reduce cognitive resources by producing task-irrelevant thinking (e.g., daydreaming) and increasing distractibility. Furthermore, boredom is proposed to induce motivation to escape from the achievement settings that cause boredom, thereby reducing students' motivation to learn. Finally, boredom is posited to impair the use of learning strategies and to promote superficial information processing instead. Given these negative effects, boredom is expected to uniformly impair students' achievement.

## Reverse Effects of Achievement on Boredom

Achievement reciprocally influences the appraisals that are thought to be proximal antecedents of boredom. More specifically, the control-value theory proposes that boredom is instigated when perceived control over achievement activities and the perceived value of these activities are too low (for empirical evidence, see Bieg, Goetz, & Hubbard, 2013; Pekrun et al., 2010). For example, if a student does not understand the contents of a lecture and perceives them as lacking relevance, perceived control and value are low, and boredom is expected to follow. Alternatively, boredom can be induced if control is too high, as implied by tasks involving low demands (e.g., repetitive monitoring tasks; Fisher, 1993). This second possibility is congruent to Csikszentmihalyi's (1975) view that boredom is caused when one's competences are high relative to task demands, which would imply high control over task performance.

Combining these propositions amounts to assuming a U-shaped curvilinear relationship between control and boredom, with boredom being promoted by either very low or very high control (i.e., a mismatch between demands and capabilities). However, some academic environments, such as the environments encountered by 1st-year students at university, pose significant challenges, making it likely that the high levels of perceived control that would promote boredom are rarely achieved in these environments. Accordingly, previous research has found that relationships between university students' perceived control and boredom were negatively linear and did not contain any curvilinear components (Pekrun et al., 2010). For the purposes of the present research, we therefore expect the relationships between students' perceived control and their boredom to be negatively linear rather than curvilinear.

Perceived control over achievement activities depends on students' individual achievement history, with success strengthening control and failure undermining it. Hence, achievement is expected to have positive effects on perceived control. By implication, since achievement has positive effects on control and control has negative effects on boredom, it follows that students' achievement should have negative effects on their academic boredom. In line with the proposed linear nature of the control–boredom link, we expect the effects of achievement on boredom to be linear as well.

## Feedback Loops of Boredom and Achievement Over Time

Because boredom is posited to influence achievement and achievement, in turn, to influence boredom, the two constructs are thought to be linked by reciprocal causation over time. Both effects are expected to be negative, amounting to positive feedback loops. This proposition implies that boredom, as a deactivating emotional experience, is different from activating negative emotions such as anxiety which may well be characterized by negative feedback loops in some individuals (e.g., failure on an exam instigating anxiety, and anxiety eliciting effort to avoid failing the next exam; Pekrun, 1992).

Feedback loops involving reduced achievement and increased boredom beg the question: What comes first, boredom or achievement? Given that achievement emotions originate early in the preschool years, this question appears to represent a chicken-and-

egg problem for later developmental phases. However, when entering a novel academic environment, boredom may develop first, triggered by diminished perceived control resulting from difficulties in understanding course material or by a lack of interest. These initial boredom experiences jeopardize students' ongoing learning behaviors, thereby negatively impacting academic achievement over time. Achievement is typically assessed later, several weeks or months into the semester. Achievement feedback then influences subsequent boredom, and feedback loops of boredom and achievement can continue to unfold across subsequent phases of studying and testing achievement.

## Overview of the Present Research

We tested the proposed reciprocal effects model using a longitudinal investigation of university students' course-related boredom and achievement. The sample included undergraduate students enrolled in introductory psychology courses spanning an entire academic year (two semesters). As noted, for testing models of reciprocal causal linkages, designs are needed that assess both variables at multiple points in time, either concurrently or in alternating order (Little, Preacher, Selig, & Card, 2007; McArdle, 2009; Rosel & Plewis, 2008). Therefore, the study used a fine-grained design that included five assessments of boredom across the academic year, as well as five assessments of performance on course exams following each boredom assessment. Keeping the design in line with our proposition that boredom can initiate feedback loops with achievement by developing early in a new environment and affecting subsequent achievement, the first assessment of boredom preceded the first exam. This study design made it possible to conduct multiple tests for the effects of boredom on subsequent performance, and of performance on subsequent boredom, while controlling for prior boredom and achievement levels.

Structural equation modeling was used to competitively test the reciprocal effects model against alternative models, including a unidirectional model only including effects of boredom on achievement (boredom effects model), a unidirectional model only including effects of achievement on boredom (achievement effects model), and an autoregressive model that did not contain any directional effects between boredom and achievement. To ensure that any observed relations were not mere artifacts of other plausible variables, we controlled for demographic and academic background variables (gender, age, and high school achievement) in each of the models. In a supplemental analysis, we additionally controlled for students' interest and intrinsic motivation to examine if the effects linking boredom and achievement were robust when including these related variables.

## Method

### Participants and Procedure

Three weeks into the academic year, 424 students were recruited from a two-semester introductory psychology course at a Canadian research-intensive university for a web-based study in exchange for experimental credit. In the initial sample, 66% were female, 79% reported English as their first language, and 89% were under 25 years of age ( $M = 20.46$  years,  $SD = 4.14$ ). Most participants



were in their 1st year of study (69%), and the average grade for students' final year of high school was 81%.

Students were required to complete a web-based questionnaire at five points throughout the first and second semesters. The first questionnaire was completed during the 3rd and 4th weeks of classes (Time 1 questionnaire). The remaining four questionnaires were completed within 10 days after each of students' next four test results were posted (Times 2–5 questionnaires). One exception was the Time 4 questionnaire that was completed during the first 10 days of the second semester due to the results of Test 3 being posted during the winter break between the two semesters. Web survey access was restricted to campus computing facilities to prevent distraction and allow for access to technical support staff, and was available only during the five time periods specified. Study reminders were provided through in-class announcements, e-mail updates, and printed notices displayed beside the posted course grades.

Some attrition occurred from one phase to the next due to students having already completed their experimental credit requirements, withdrawing from the course, or illness. However, the extent of the attrition observed was minimal: 4% from Times 1 to 2, 5% from Times 2 to 3, 6% from Times 3 to 4, and 1% from Times 4 to 5; total attrition rate was 16%. These attrition rates show considerable engagement in the web-based study protocol and are below those observed in similar pencil-and-paper studies (e.g., 21%: Hall, Perry, Chipperfield, Clifton, & Haynes, 2006; 20%: Perry, Hladkyj, Pekrun, & Pelletier, 2001).

A regression analysis on a continuous measure of study attrition was conducted (number of boredom assessments not completed;  $M = 0.43$ ,  $SD = 1.07$ , range = 0–4). Predictors included gender, age, high-school grades, initial course performance (Test 1), and Time 1 boredom. Results showed students with better initial performance to withdraw from fewer assessments ( $\beta = -.27$ ,  $p < .01$ ). However, the proportion of variance in study attrition explained by the predictors was small ( $R^2 = .10$ ).

## Study Measures

The measures included a self-report scale of boredom, objective test performance, demographic background variables (gender, age, and high-school grades), and affective background variables (interest and intrinsic motivation). Means, standard deviations, and ranges for all measures in each study phase are outlined in Table 1.

**Boredom.** A six-item version of the learning-related boredom scale of the Achievement Emotions Questionnaire (AEQ; Pekrun et al., 2011) was used to assess students' learning-related boredom concerning their psychology course (e.g., "When studying for this course, I feel bored"; 1 = *not at all true*, 5 = *completely true*; see Appendix A for the scale items). The measure showed high internal reliability ( $\alpha$ s = .88, .89, .90, .91, and .92, for Times 1, 2, 3, 4, and 5, respectively).

**Achievement.** Students' grade percentages on the five tests in introductory psychology that followed the boredom assessments were obtained from course instructors throughout the academic year. Each exam was criterion-referenced, of equal weight, non-cumulative in content, and involved a multiple-choice format with equivalent numbers of items. The tests were administered approximately one month apart.

### Covariates.

**Demographic background variables.** Three demographic background variables were included in the analyses—namely, gender, age, and high-school grades. High-school grades consisted of students' average final grade, computed as a percentage, in university pre-requisite courses (i.e., English, mathematics, chemistry, physics) completed during their final year of high school. Since Scholastic Aptitude Tests (SATs) are not administered to Canadian students, high-school grades were used as a proxy for pre-existing aptitude differences, based on research showing high-school achievement to strongly predict college performance (e.g., Hoffman, 2002; Zheng, Saunders, Shelley, & Whalen, 2002).

**Interest and intrinsic motivation.** Measures of interest and intrinsic motivation were available in the Time 1 questionnaire and

Table 1  
*Descriptive Statistics for the Study Variables*

Variable	<i>M</i>	<i>SD</i>	Possible range	Observed range
Boredom				
Time 1	11.40	4.21	6–30	6–30
Time 2	11.38	4.17	6–30	6–29
Time 3	11.77	4.62	6–30	6–30
Time 4	12.51	5.13	6–30	6–30
Time 5	12.70	5.21	6–30	6–30
Performance				
Test 1	78.26	13.14	0–100	22.81–100
Test 2	73.54	14.12	0–100	27.78–100
Test 3	72.90	14.01	0–100	18.46–100
Test 4	74.44	14.54	0–100	13.20–100
Test 5	73.98	14.23	0–100	26.09–100
Demographic background variables				
Age	20.46	4.14		17–45
High-school grades	81.06	8.70	0–99	55–98
Affective background variables				
Interest	8.38	1.55	1–10	2–10
Intrinsic motivation	24.00	4.33	5–35	9–33

used to control for these constructs in the supplemental analysis. Students' interest in the course was measured using one self-report item ("I think that what we learn in my Introductory Psychology course is interesting"; 1 = *strongly disagree*, 10 = *strongly agree*). Intrinsic motivation was measured based on the intrinsic goal orientation scale from the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich, Smith, Garcia, & McKeachie, 1991; five items, e.g., "I prefer course material that arouses my curiosity, even if it is difficult to learn"; 1 = *not at all true of me*, 7 = *very true of me*;  $\alpha = .71$ ).

### Rationale for Structural Equation Modeling

Structural equation modeling (Mplus, Version 7; Muthén & Muthén, 2012) was used to evaluate the reciprocal effects model and test it against more constrained models. The model represents a sequential analysis of reciprocal effects consistent with the sequential manner in which the measures were assessed (for a similar procedure, see, e.g., Marsh & O'Mara, 2008). In contrast to a traditional cross-lagged model in which variables are assessed simultaneously within each measurement occasion, boredom and performance were modeled in alternating order consistent with the data collection process (see Figure 1). As such, the present model includes five paths from boredom to performance and four paths from performance to boredom. The five boredom variables were modeled as latent constructs. The five test performance measures and the three background measures (gender, age, and high-school grades) were evaluated as manifest variables. The background variables were included as covariates; for each of these variables, directional paths to all of the boredom and performance variables were included, as were correlations between the background variables.

**Measurement model for boredom.** The six boredom scale items were used as indicators for each of the five latent boredom variables. Following recommendations by Pekrun et al. (2011), a correlated uniquenesses approach was used to model boredom within each boredom assessment by including correlations between residuals for items representing the same emotion component (Items 1 and 2, 3 and 4, and 5 and 6 for the affective, cognitive, and physiological-motivational components of bore-

dom, respectively). In addition, correlations between residuals for the boredom items across measurement occasions were included to control for systematic measurement error.

**Hierarchical data structure, estimator used, and missing values.** The university course from which the study sample was drawn consisted of five sections taught by different instructors. As students were nested in these sections, we corrected for the clustering of the data using the "type = complex" option implemented in Mplus (Muthén & Muthén, 2012). To estimate the model parameters, the robust maximum likelihood estimator (MLR) was employed, which is robust to nonnormality of the observed variables. In order to make full use of the data from students who had missing data, we applied the full information maximum likelihood method (FIML; Enders, 2006) implemented in Mplus.

**Sequential testing of the reciprocal effects model.** The reciprocal effects model was tested in a sequential manner. We first tested the measurement invariance of the boredom measure across time by comparing two measurement models, an unconstrained baseline model and a strict factorial invariance model that constrained factor loadings, item intercepts, and item residuals to be equal across the five measurement occasions (Brown, 2006). Subsequently, we tested the reciprocal effects model (Model 1; Figure 1) against three alternative models (see Figure 2; McArdle, 2009): a boredom effects model that estimated effects of boredom on subsequent achievement but constrained the effects of achievement on boredom to be zero (Model 2); an achievement effects model that estimated the effects of achievement on boredom but constrained the effects of boredom on achievement to be zero (Model 3); and an autoregressive model that constrained all effects of boredom on performance, and vice versa, to be zero (Model 4). Autoregressive effects for boredom and achievement were included in all of these models, and the background variables were included as covariates. Finally, we conducted a supplemental analysis of the reciprocal effects model that additionally included interest and intrinsic motivation as covariates (Model 5). In this model, interest was evaluated as a manifest variable, and intrinsic motivation as a latent variable using the five items of the intrinsic motivation scale as manifest indicators.

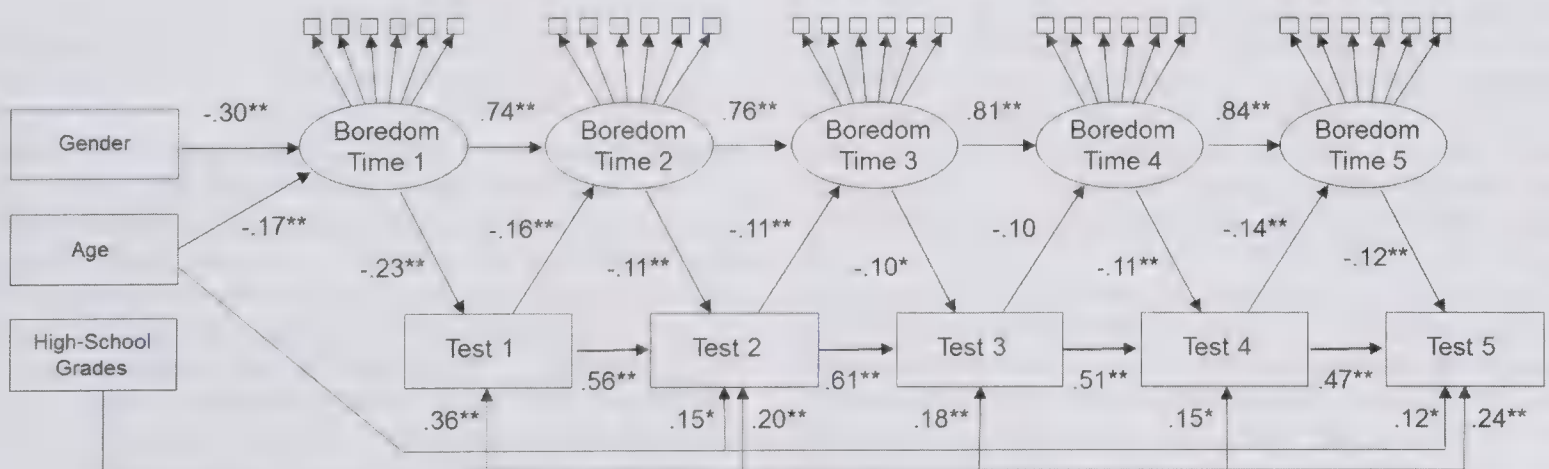


Figure 1. Results for Model 1 (reciprocal effects model). For the covariates, significant paths are displayed only (see Table 3 for more complete information). Gender was coded male = 1, female = 2. \* $p < .05$ . \*\* $p < .01$ .



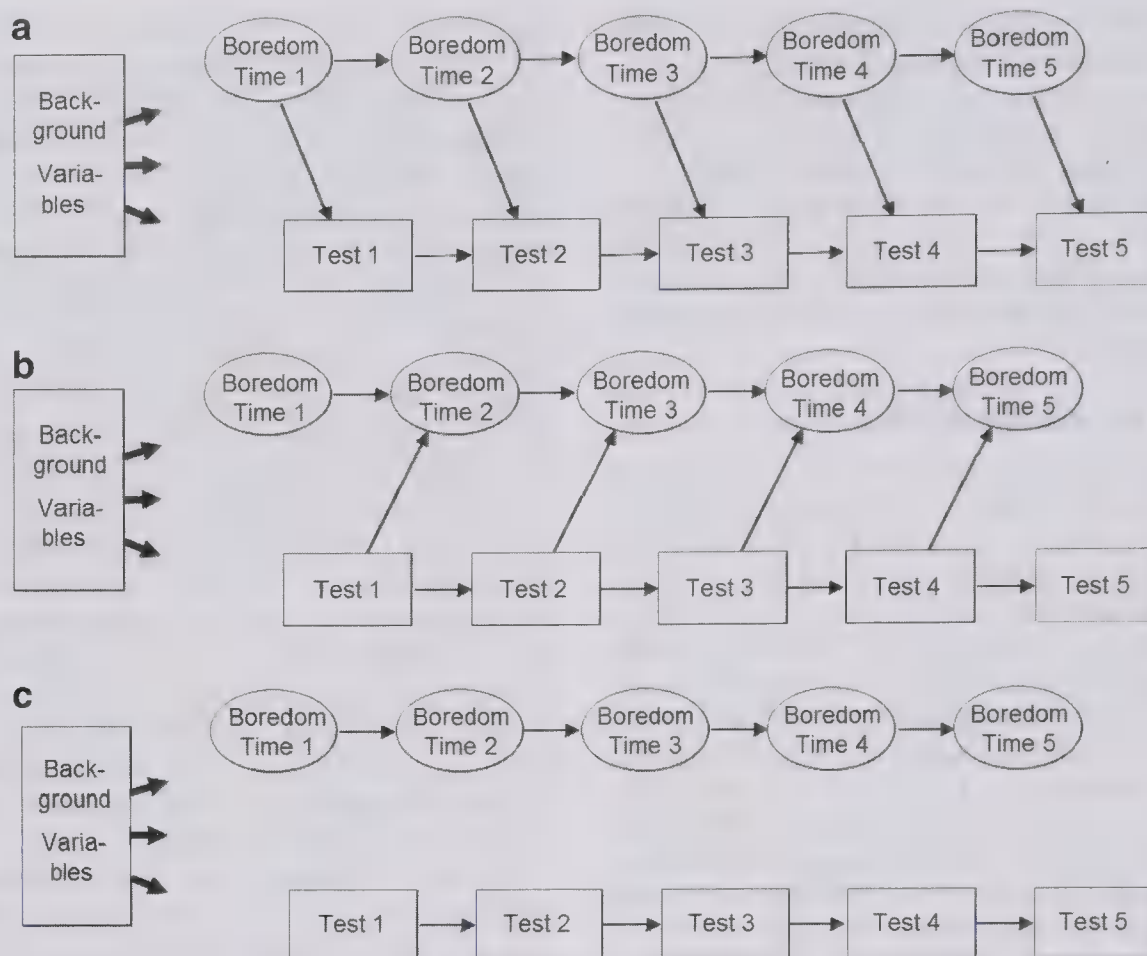


Figure 2. Structure of Models 2–4. (a) Model 2 (boredom effects model). (b) Model 3 (achievement effects model). (c) Model 4 (autoregressive model).

**Goodness-of-fit indexes to evaluate model fit.** We applied both absolute and incremental fit indices to evaluate the fit of the models, including the comparative fit index (CFI), the Tucker–Lewis index (TLI), the root-mean-square error of approximation (RMSEA), and the standardized root-mean-square residual (SRMR). Traditionally, values of CFI and TLI close to .95, values of RMSEA lower than .06, and values of SRMR lower than .08 have been interpreted as indicating good fit (Browne & Cudeck, 1993; Hu & Bentler, 1998; MacCallum, Browne, & Sugawara, 1996). Following the rationale used by Trautwein et al. (2012), we considered a model to have reasonably good fit to the observed data when at least two of these criteria were fulfilled. However, it should be noted that these recommended cutoff values were originally derived from analyses with relatively simple simulated data sets. These values are often not met with data sets derived from more complex studies, suggesting that they should be used with caution (Heene, Hilbert, Draxler, Ziegler, & Bühner, 2011; Marsh, Hau, & Wen, 2004).

For comparing nested models (i.e., Models 1–4 and the unconstrained vs. factorial invariance models for the latent boredom variable), we used the Satorra–Bentler scaled chi-square difference test including scaling corrections for nestedness (Bryant & Satorra, 2012; Satorra, 2000), which is suited for use with the MLR estimator. This test provides the chi-square difference statistic TRd that is corrected for nonnormality of the observed variables and nestedness of the models. In addition, we evaluated the Akaike information criterion (AIC; Akaike, 1974) and the sample-size

corrected Bayesian information criterion (BIC; Schwarz, 1978). Lower values of these criteria indicate better model fit. Using these criteria, the model with the smallest AIC and BIC should be chosen. For interpreting values of AIC and BIC, it is important to note that it is not the absolute size of the values but the difference between values which is relevant. For AIC, differences of  $\Delta AIC > 10$  are considered as substantial and indicating that the model obtaining the higher value has essentially no empirical support (Burnham & Anderson, 2002).

## Results

### Preliminary Analyses

**Correlations.** Correlations between boredom, interest, intrinsic motivation, performance, and the background variables are outlined in Table 2. Correlations between the boredom measures ( $r$  range = .59–.81) and between the performance outcomes ( $r$  range = .56–.71) over time indicated a substantial degree of stability for both variables, with the highest correlations found between adjacent assessments. Furthermore, consistently negative correlations between boredom and performance were observed ( $r$  range = –.22 to –.36). Interest and intrinsic motivation correlated negatively with boredom, and interest correlated positively with performance on Test 1 and Test 2. Concerning the academic background variables, high-school grades were positively correlated with course performance ( $r$  range = .33–.41).

Table 2  
Pearson Product Moment Correlations for the Study Variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Boredom—Time 1	—														
2. Boredom—Time 2	.70**	—													
3. Boredom—Time 3	.64**	.69**	—												
4. Boredom—Time 4	.62**	.71**	.76**	—											
5. Boredom—Time 5	.59**	.71**	.76**	.81**	—										
6. Test 1	-.23**	-.33**	-.32**	-.25**	-.30**	—									
7. Test 2	-.30**	-.35**	-.33**	-.31**	-.31**	.68**	—								
8. Test 3	-.22**	-.30**	-.31**	-.35**	-.33**	.63**	.71**	—							
9. Test 4	-.26**	-.25**	-.28**	-.29**	-.36**	.61**	.62**	.60**	—						
10. Test 5	-.23**	-.30**	-.29**	-.32**	-.31**	.56**	.65**	.70**	.61**	—					
11. Gender	-.27**	-.24**	-.21**	-.21**	-.18**	.08	.09	.12*	.08	.16**	—				
12. Age	-.14**	-.10	-.11*	-.13*	-.20**	-.02	.10*	.08	.13**	.15**	-.05	—			
13. High-school grades	.00	-.07	-.06	-.03	.00	.34**	.38**	.41**	.33	.36**	.12*	-.25**	—		
14. Interest	-.53**	-.48**	-.45**	-.38**	-.38**	.11*	.15**	.09	.12	.07	.13*	.13*	-.08	—	
15. Intrinsic motivation	-.32**	-.26**	-.26**	-.25**	-.25**	.01	.06	.03	.08	.05	.01	.13**	.00	.47**	—

\*  $p < .05$ . \*\*  $p < .01$ .

**Mean level change and gender effects on boredom.** A repeated-measures ANOVA was conducted to assess boredom as a function of time and gender throughout the academic year. A significant main effect of time was found,  $F(4, 1216) = 16.19$ ,  $p < .01$ ,  $\eta_p^2 = .05$ , showing boredom levels to increase over time (means are provided in Table 1). Gender also had a main effect on boredom, with males ( $M = 13.29$ ,  $SD = 4.06$ ) reporting greater boredom overall relative to females ( $M = 11.40$ ,  $SD = 4.11$ ),  $F(1, 304) = 13.58$ ,  $p < .01$ ,  $\eta_p^2 = .04$ .

**Test of linearity for the achievement–boredom link.** The propositions of the control-value theory (Pekrun, 2006) imply that the effects of achievement on boredom can take curvilinear forms, as noted earlier. For academic settings at university, however, we expected that these effects would be negatively linear. To test for linearity, we performed simultaneous multiple regression analysis for the links between test performance and subsequent boredom scores. The regression equations included linear and quadratic terms for performance which were computed after centering the performance variable. This was done separately for each test assessment that was followed by a boredom assessment, thus involving four analyses (for the effects of Tests 1, 2, 3, and 4 on Boredom Times 2, 3, 4, and 5, respectively). Across measurement occasions, test performance had a significant linear effect on subsequent boredom ( $\beta_s = -.33$ ,  $-.31$ ,  $-.34$ , and  $-.42$  for the effects of Tests 1, 2, 3, and 4 on subsequent boredom; all  $ps < .01$ ).

There were no significant effects for the quadratic term in any of the equations, with one exception. In the Test 4/Boredom Time 5 analysis, the effect of the quadratic term was significant ( $\beta = -.15$ ,  $p < .01$ ; unstandardized  $B = -.002$ ). However, this effect was small relative to the effect of the linear term. An inspection of the regression function showed that the effect implied a decrease of the negative slope of the regression curve with increased performance (i.e., a slowing down of the decrease of boredom). There were no positive effects of performance on boredom at any interval of the regression curve. As such, the Test 4/Boredom Time 5 regression function was monotonically negative and nearly linear. Overall, these findings show that achievement had negative effects on subsequent boredom that were linear, or approximately linear,

for each of the study phases, corroborating our hypothesis that the effects of university students' achievement on their boredom are simply negative and do not follow a U-shaped function.

## Results of Structural Equation Modeling

**Measurement invariance of the boredom scale.** Confirmatory factor analysis was used to evaluate the measurement equivalence of the boredom measure across the five boredom assessments. We evaluated the strict factorial invariance model (Brown, 2006; Meredith, 1993) that provides a strong test of equivalence by constraining factor loadings (metric invariance), item intercepts (scalar invariance), and item error variances (invariant uniquenesses) to be equal across measurement occasions. The model showed a good fit to the data,  $\chi^2(384) = 559.37$ ,  $p < .01$ ; CFI = .975; TLI = .971; RMSEA = .033; SRMR = .040. The fit of an unconstrained baseline model that allowed parameters to vary across time was as follows:  $\chi^2(320) = 478.85$ ,  $p < .01$ ; CFI = .977; TLI = .969; RMSEA = .034; SRMR = .033. Comparing the two models, the Satorra–Bentler scaled chi-square difference test including scaling corrections for nestedness was not significant (TRd [64] = 81.78,  $p > .05$ ). Moreover, the AIC and the sample-size corrected BIC were lower for the strict factorial invariance model (AIC = 22,586.41 and 22,556.75, and BIC = 22,737.72 and 22,652.72, for the unconstrained and factorial invariance models, respectively;  $\Delta AIC = 29.66$ ). This finding suggests that the strict factorial invariance model is preferable to the unconstrained model when using an information-theoretical perspective. Overall, these results clearly indicate that the strict factorial invariance model could be accepted, thus documenting that the boredom measure exhibited measurement invariance over time.

**Reciprocal effects model (Model 1).** The reciprocal effects model included effects of boredom on performance and reverse effects of performance on boredom across all assessments of boredom and performance, as well as autoregressive effects and effects of the covariates (gender, age, and high school grades) on all of the boredom and performance variables (see Figure 1 and Table 3). The fit indices provided good support for this model,  $\chi^2(548) = 1,136.05$ ,  $p < .01$ ; CFI = .931; TLI = .912; RMSEA =



Table 3

*Standardized Factor Loadings, Path Coefficients, and Residual Variances for Model 1 (Reciprocal Effects Model)*

Coefficient	Boredom					Test performance				
	Time 1	Time 2	Time 3	Time 4	Time 5	Test 1	Test 2	Test 3	Test 4	Test 5
Factor loadings										
Item 1	.76**	.72**	.82**	.79**	.85**					
Item 2	.77**	.84**	.81**	.83**	.83**					
Item 3	.86**	.87**	.88**	.90**	.87**					
Item 4	.61**	.63**	.66**	.74**	.74**					
Item 5	.70**	.79**	.77**	.80**	.82**					
Item 6	.76**	.72**	.72**	.73**	.78**					
Path coefficients										
Boredom Time $n^a$						-.23**	-.11*	-.10*	-.11**	-.12**
Boredom Time $n - 1^b$		.74**	.76**	.81**	.84**					
Test $n - 1^c$		-.16**	-.11**	-.10	-.14**		.58**	.62**	.52**	.47**
Gender	-.30**	-.05	-.04	-.05	.00	-.03	.01	.03	-.01	.08
Age	-.17**	.00	-.02	-.03	-.04	.03	.15**	.01	.11	.12*
High-School grades	.02	.00	.02	.05	.04	.36**	.20**	.18**	.15*	.24**
Residual variances	.89**	.34**	.34**	.25**	.20**	.83**	.48**	.44**	.59**	.54**

<sup>a</sup> Effects of Boredom Times 1, 2, 3, 4, and 5 on Tests 1, 2, 3, 4, and 5, respectively. <sup>b</sup> Effects of Boredom Times 1, 2, 3, and 4 on Boredom Times 2, 3, 4, and 5, respectively. <sup>c</sup> Effects of Tests 1, 2, 3, and 4 on Boredom Times 2, 3, 4, and 5 and Tests 2, 3, 4, and 5, respectively.

\*  $p < .05$ . \*\*  $p < .01$ .

.050; SRMR = .063. The AIC and samples-size corrected BIC were 43,843.38 and 44,045.83, respectively. Both boredom and performance showed considerable stability over time, with autoregressive coefficients for boredom in the  $\beta = .74$ –.84 range, and for exam scores in the  $\beta = .47$ –.61 range. Regarding the covariates, gender and age negatively predicted boredom at Time 1, and high-school grades positively predicted exam performance on Tests 1–5.

Despite considerable stability of the performance variable and substantial effects of prior achievement on performance, results showed boredom to negatively predict each subsequent performance outcome, with the strongest path observed between the initial boredom and performance variables ( $\beta$  range =  $-.10$  to  $-.23$ ; see Figure 1). Negative paths from each test outcome to the subsequent boredom variable were also observed ( $\beta$  range =  $-.10$  to  $-.16$ ), with three out of the four path coefficients being significant. The effect of Test 3 on Boredom Time 4 ( $\beta = -.10$ , *ns*) represented the influence of the last exam before the winter break on boredom assessed after the winter break; the non-significance of the effect may have been due to the time lag and intervening events during this break. As such, the results provide empirical support for the study hypotheses in showing a consistent sequence of negative paths from boredom to performance, and vice versa, throughout the academic year.

**Comparison with the boredom effects, achievement effects and autoregressive models (Models 2–4).** The unidirectional boredom effects and achievement effects models had the same structure as the reciprocal effects model, but some of the effects linking boredom and achievement were constrained to be zero. Specifically, the boredom effects model included effects of boredom on achievement, and the achievement effects model included effects of achievement on boredom, with reverse effects being constrained to zero in these models (see Figure 2). Fit indices for the boredom effects model were as follows:  $\chi^2(552) = 1,189.20$ ,  $p < .01$ ; CFI = .926; TLI = .906; RMSEA = .052; SRMR = .096; AIC = 43,879.20; sample-size corrected BIC = 44,078.14.

Fit indexes for the achievement effects model were as follows:  $\chi^2(552) = 1,183.90$ ,  $p < .01$ ; CFI = .926; TLI = .906; RMSEA = .052; SRMR = .076; AIC = 43,881.34; sample-size corrected BIC = 44,080.28. Both of these models fit the data significantly worse than the reciprocal effects model, with TRd (4) = 104.43,  $p < .01$ , for the boredom effects model, and TRd (4) = 60.73,  $p < .01$ , for the achievement effects model. In addition, AIC and BIC were substantially higher than for the reciprocal effects model ( $\Delta$ AIC = 35.82 and 37.96 for the boredom effects and achievement effects models, respectively). These findings clearly indicate that the reciprocal effects model is preferable to these two unidirectional models.

In the autoregressive model, all of the effects linking boredom and achievement were constrained to be zero. Fit indexes for this model were as follows:  $\chi^2(556) = 1,239.37$ ,  $p < .01$ ; CFI = .920; TLI = .900; RMSEA = .054; SRMR = .118; AIC = 43,925.78; sample-size corrected BIC = 44,121.21. The Satorra–Bentler scaled chi-square difference test showed that the model fit the data less well than the reciprocal effects model, TRd (8) = 140.68,  $p < .01$ . In addition, AIC and BIC were substantially higher than for the reciprocal effects model ( $\Delta$ AIC = 82.40), which also suggests poorer fit for the autoregressive model. These findings indicate that the reciprocal effects model is superior to the autoregressive model. In sum, the findings clearly indicate that the reciprocal effects model fit the data significantly better, and can be judged to be more likely given the data (see Burnham & Anderson, 2002, Chapter 2), as compared with any of three alternative models.

**Supplemental analysis: Controlling for interest and intrinsic motivation (Model 5).** In a supplemental analysis, we expanded the reciprocal effects model by additionally controlling for students' interest and intrinsic motivation assessed at Time 1. By testing whether the links between boredom and achievement were sufficiently robust when these related variables were included, we sought to address a potential concern that the boredom construct may simply be regarded as the inverse of interest or intrinsic motivation, as discussed at the outset. The fit indexes for this

model were similar to the indexes for the original Model 1, with  $\chi^2(759) = 1,532.81$ ,  $p < .01$ ; CFI = .919; TLI = .900; RMSEA = .049; SRMR = .063. Again, gender and age had negative effects on boredom at Time 1, and high-school grades had positive effects on performance on all of the course exams. Interest also had negative effects on boredom. Furthermore, interest had a positive effect on performance on Test 1 (see Appendix B for the estimated model parameters).

Of critical importance, the path coefficients for the effects linking boredom and achievement replicated the coefficients of the original model. Again, all the effects of boredom on achievement were significantly negative, with  $\beta$ s ranging from  $-.09$  to  $-.25$ . The effects of achievement on boredom were negative as well, with  $\beta$ s ranging from  $-.11$  to  $-.15$  and three out of the four coefficients being significant. Again, there was one non-significant effect (Test 3 on Boredom Time 4;  $\beta = -.11$ , *ns*) which pertained to the effect of test performance on students' boredom across the winter break. In sum, these results show that the effects of boredom on achievement, and the effects of achievement on boredom, were robust when controlling for students' interest in the course and their intrinsic motivation.

## Discussion

The findings of this study provide evidence for the proposed reciprocal effects model of boredom and achievement. As suggested by longitudinal structural equation modeling, university students' course-related boredom had negative effects on their performance on subsequent course exams, and exam performance, in turn, had negative effects on subsequent boredom. The findings were consistent across all assessments of boredom and performance, except the link between the last exam taken prior to the winter break and boredom after the break which was negative but not significant. Alternative models including only effects of boredom on performance, or only effects of performance on boredom, also showed a reasonable fit to the data in terms of CFI and RMSEA; however, as indicated by chi-square difference tests and the comparison of AIC values, the reciprocal effects model clearly showed better fit than these alternative models.

Because prior boredom and achievement as well as demographic background variables were controlled, the path coefficients are likely to represent effects of boredom on achievement, and vice versa, rather than simply the influence of prior boredom, prior achievement, gender, or age. This was further supported by supplemental findings showing the boredom–performance link to be robust when additionally controlling for students' interest and intrinsic motivation.

For interpreting the size of the path coefficients linking boredom and performance, it is important to note that the coefficients represent incremental effects due to prior boredom and achievement being controlled. Thus, the coefficients represent effects of each variable on change in the other from one assessment to the next, rather than effects on the absolute levels of these variables. Furthermore, both boredom and performance showed considerable stability over time, leaving little variance to be explained and making it difficult to detect the effects of additional variables. From this perspective, the consistency of effects lends credibility to the notion that boredom and achievement are indeed linked by reciprocal causation over time.

## Reciprocal Effects of Boredom and Achievement

The present findings add to the research literature by documenting that boredom negatively predicts students' scholastic attainment over time. This is congruent with previous evidence showing that boredom and academic achievement are negatively correlated, as summarized at the outset. However, the present findings go beyond correlational evidence by disentangling the directional effects underlying the boredom–achievement link. Specifically, the findings indicate that boredom indeed has a negative influence on students' achievement, over and above the effects of prior accomplishments. These negative effects are in line with propositions derived from Pekrun's (2006) control-value theory, which posits that boredom has uniformly negative effects on learning and achievement outcomes.

The results also contribute to our understanding of the origins of students' boredom. The findings indicate that achievement, in turn, had negative effects on boredom, implying that successful completion of exams can reduce students' boredom, whereas doing poorly exacerbates their boredom. These effects are likely mediated by students' perceptions of control over achievement, with low control leading to greater boredom (Pekrun et al., 2010). The findings of regression analyses further showed the links between achievement and subsequent boredom to be virtually linear in nature rather than representing U-shaped curvilinear effects.

These negative effects of performance on boredom are counter to the accepted view that boredom is primarily experienced by gifted students who are not sufficiently challenged by academic demands. However, they are consistent with our earlier reasoning that university courses pose considerable challenges for many students, so that even the most capable students must struggle to retain perceived control and to master the material, contrary to the notion that success at university comes easy, due to boring and routine task demands. This may be especially true for 1st-year students, who are the focus of the present study.

Taken together, these negative effects amount to positive feedback loops linking the two constructs. In a few longitudinal studies, previous research has found students' test anxiety and academic achievement to be linked by positive feedback loops (Meece, Wigfield, & Eccles, 1990; Pekrun, 1992). The present research adds to this literature by showing that boredom, an underexplored academic emotion, demonstrates similar links with performance. As such, it seems that unidirectional models cannot adequately capture the complex reality of students' emotions. Rather, systems-oriented perspectives (Turner & Waugh, 2007) are needed that take more complex patterns of causal links into account, including feedback loops between emotions, their antecedents, and their effects.

## Effects of Interest and Intrinsic Motivation

In our supplemental analyses, interest had negative effects on students' boredom. The effects of intrinsic motivation on boredom were also negative, although most of them were not significant. Although a one-item measure cannot substitute for a more comprehensive assessment of interest, the findings suggest that interest can protect against feeling bored, and that lack of interest can contribute to the arousal of boredom (Pekrun et al., 2010). In addition, interest had a positive effect on initial exam performance. The small size of this effect and the lack of positive performance



effects of intrinsic motivation are in line with previous research showing that interest and intrinsic motivation typically do not have a strong influence on students' immediate academic performance; rather, they influence students' long-term attainment and academic choices (Murayama, Pekrun, Lichtenfeld, & vom Hofe, 2013; Schiefele, 2009). To the extent that this is true, boredom and interest (or lack of interest) may have asymmetrical effects, with boredom immediately impacting learning and interest primarily having long-term effects.

### Development of Boredom Over Time

Students' boredom was found to increase over the academic year. This increase is equivalent with the development of boredom observed for middle and high school students (Ahmed et al., 2013; Pekrun et al., 2007), and with the decline of interest and intrinsic motivation that is observed during adolescence after students have entered middle school (e.g., Fredricks & Eccles, 2002; Frenzel, Goetz, Pekrun, & Watt, 2010). To the extent that there are common mechanisms of affective change during young adulthood and adolescence, one explanation may be that boredom goes up, and enjoyment down, after the initial excitement experienced in a new educational environment has dissipated.

A second possible explanation is provided by our reciprocal effects model. The feedback loops predicted by the model imply that there should be a symmetrical, self-sustaining development of boredom and achievement over time, with initial boredom reducing subsequent achievement, and reduced achievement contributing to increased boredom. Multiple cycles of this type should lead to a steady increase of average boredom scores across time, as observed in the present research. They should also lead to a reduction of achievement over time; however, the present analysis is not suited to examine this prediction, as exam scores at university lack a common metric due to variation in the contents and difficulty of exams across the academic year.

### Limitations, Suggestions for Future Research, and Implications for Practice

The present investigation represents a significant advance over previous research by documenting reciprocal effects of boredom and achievement over time, while controlling for critical affective and demographic background variables. Nevertheless, several limitations should be considered when interpreting the study findings, and can be used to suggest directions for future research.

**Methodological considerations.** The power of non-experimental field studies to derive causal conclusions is limited by the nature of their design. Clearly, non-experimental designs are less powerful for deriving causal conclusions than experimental designs, all other things being equal. As such, although the present analysis used multi-wave longitudinal structural equation modeling and controlled for related variables and autoregressive effects, it cannot be completely ruled out that the study findings were due to other variables not included in the study. On the other hand, field studies of emotion may have more ecological validity than experimental studies that typically are limited in terms of situational representativeness (partially due to ethical limits on experimentally manipulating emotions): Emotion research takes no exception regarding the trade-off between internal and external validity that is typical of scientific inquiry in psychology.

As such, the power of non-experimental field studies to derive conclusions regarding real-world causal processes in emotion may be limited due to threats to internal validity, whereas the power of experimental studies to derive such conclusions may be limited in terms of reduced external validity. By implication, future research should further pursue the approach taken herein but should also complement this approach with experimental studies on the link between boredom and students' achievement.

One specific methodological limitation of the present analysis is that achievement was modeled as a manifest variable. By using exam grades, we sought to employ an ecologically valid measure of student achievement. As is typical for grades, information about reliability was not available; as such, it was not possible to disattenuate the boredom-achievement link for potential unreliability of the achievement measure. From the perspective of hypothesis testing, this implies that our study hypotheses were tested in a conservative manner. With additional correction for measurement error in the achievement variables, the effects of boredom on achievement may have been even stronger than in the present analysis.

However, there may also be an alternative perspective on the reliability of grades. As a measure of achievement, grades may have less than perfect reliability, implying that any effects of boredom on achievement may be underestimated. By contrast, from the perspective of grades as sources of students' affective development, they could be seen as having almost perfect reliability, as grades rather than true achievement provide the feedback that shapes students' perceptions of success and failure.

**Substantive issues.** The present research examined academic boredom as experienced by university students. Compared with the general population, university students represent a select group having above-average ability and positive achievement experiences throughout prior stages of the educational career. It is open to question whether the present findings would generalize to low-ability adults, to younger age groups, and to the general student population in K–12 educational institutions. As the present research involved samples of North American students, it also remains an open question as to whether the findings would generalize to students in other cultures.

Regarding the origins of students' boredom, the present findings indicate that poor academic achievement can contribute to the arousal of boredom. As such, the findings suggest that poor achievement and academic demands that are too challenging can trigger boredom in university students. Future research should examine under which task conditions, and in which students, boredom may also occur due to academic demands that are too low and fail to challenge students' competencies, in line with Csikszentmihalyi's (1975) suggestions concerning the impact of low task demands (also see Daschmann, Goetz, & Stupnisky, 2011; Nett, Goetz, & Hall, 2011). To this end, it may be useful to examine boredom across various academic contexts and in students representing widely differing levels of ability, including gifted students who may more easily experience high levels of control over academic demands.

Finally, the study addressed the overall relation between boredom and achievement but did not examine the mechanisms that may mediate the observed links. In the proposed model of reciprocal effects, it is posited that effects of boredom on achievement are due to the detrimental influence of boredom on cognitive



resources, motivation, strategy use, and self-regulation, and that the effects of achievement outcomes on boredom are mediated by perceptions of control over performance. More research on the link between boredom and achievement as mediated by cognitive and motivational mechanisms is needed to better understand students' boredom and to inform efforts to remediate its deleterious effects.

**Implications for educational practice.** Two important messages can be derived from the present research. First, the study results suggest that boredom has uniformly negative effects on students' academic achievement, and these effects are not mere epiphenomena of prior performance: More likely, they represent a true negative causal influence of students' boredom experiences. By implication, the findings suggest that educators, administrators, and parents alike may want to consider intensifying efforts that minimize students' boredom, the relative inconspicuousness of this emotion notwithstanding. Second, the results imply that achievement outcomes reciprocally influence students' boredom, suggesting that successful performance attainment and positive achievement feedback can contribute to a reduction of students' boredom, whereas failure experiences can increase boredom (also see Pekrun, Cusack, Murayama, Elliot, & Thomas, 2014). Accordingly, providing students with experiences of success (e.g., in terms of mastery learning) may help to prevent the development of boredom. By documenting the influence of achievement outcomes on students' boredom, the present findings elucidate one important factor that can be targeted by educators to reduce negative affect and thereby facilitate students' academic development.

## References

- Ahmed, W., van der Werf, G., Kuyper, H., & Minnaert, A. (2013). Emotions, self-regulated learning, and achievement in mathematics: A growth curve analysis. *Journal of Educational Psychology, 105*, 150–161. doi:10.1037/a0030160
- Ainley, M., & Hidi, S. (in press). Interest and enjoyment. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *Handbook of emotions in education*. New York, NY: Taylor & Francis.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723. doi:10.1109/TAC.1974.1100705
- Amos, A., Wiltshire, S., Haw, S., & McNeill, A. (2006). Ambivalence and uncertainty: Experiences of and attitudes toward addiction and smoking cessation in the mid-to-late teens. *Health Education Research, 21*, 181–191. doi:10.1093/her/cyh054
- Bieg, M., Goetz, T., & Hubbard, K. (2013). Can I master it and does it matter? An intraindividual analysis on control-value antecedents of trait and state academic emotions. *Learning and Individual Differences, 28*, 102–108. doi:10.1016/j.lindif.2013.09.006
- Blaszczynski, A., McConaghy, N., & Frankova, A. (1990). Boredom proneness in psychopathological gambling. *Psychological Reports, 67*, 35–42.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–161). Thousand Oaks, CA: Sage.
- Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling, 19*, 372–398. doi:10.1080/10705511.2012.687671
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inferences: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer.
- Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety*. San Francisco, CA: Jossey-Bass.
- Daniels, L. M., Stupnisky, R. H., Pekrun, R., Haynes, T. L., Perry, R. P., & Newall, N. E. (2009). A longitudinal analysis of achievement goals: From affective antecedents to emotional effects and achievement outcomes. *Journal of Educational Psychology, 101*, 948–963. doi:10.1037/a0016096
- Daschmann, E. C., Goetz, T., & Stupnisky, R. H. (2011). Testing the predictors of boredom at school: Development and validation of the Precursors to Boredom Scales. *British Journal of Educational Psychology, 81*, 421–440. doi:10.1348/000709910X526038
- Enders, C. K. (2006). A primer on the use of modern missing-data methods in psychosomatic medicine research. *Psychosomatic Medicine, 68*, 427–436. doi:10.1097/01.psy.0000221275.75056.d8
- Farmer, R., & Sundberg, N. D. (1986). Boredom proneness—The development and correlates of a new scale. *Journal of Personality Assessment, 50*, 4–17. doi:10.1207/s15327752jpa5001\_2
- Fisher, C. D. (1993). Boredom at work: A neglected concept. *Human Relations, 46*, 395–417. doi:10.1177/001872679304600305
- Fogelman, K. (1976). Bored eleven-year-olds. *British Journal of Social Work, 6*, 201–211.
- Folkman, S., & Lazarus, R. S. (1985). If it changes it must be a process: Study of emotion and coping during three stages of a college examination. *Journal of Personality and Social Psychology, 48*, 150–170. doi:10.1037/0022-3514.48.1.150
- Fredricks, J. A., & Eccles, J. (2002). Children's competence and value beliefs from childhood through adolescence: Growth trajectories in two male-sex-typed domains. *Developmental Psychology, 38*, 519–533. doi:10.1037/0012-1649.38.4.519
- Frenzel, A. C., Goetz, T., Pekrun, R., & Watt, H. M. G. (2010). Development of mathematics interest in adolescence: Influences of gender, family, and school context. *Journal of Research on Adolescence, 20*, 507–537. doi:10.1111/j.1532-7795.2010.00645.x
- Frenzel, A. C., Pekrun, R., & Goetz, T. (2007). Perceived learning environments and students' emotional experiences: A multilevel analysis of mathematics classrooms. *Learning and Instruction, 17*, 478–493. doi:10.1016/j.learninstruc.2007.09.001
- Goetz, T., Frenzel, A. C., Hall, N. C., Nett, U., Pekrun, R., & Lipnevich, A. (in press). Types of boredom: An experience sampling approach. *Motivation and Emotion*.
- Goetz, T., Frenzel, A. C., Pekrun, R., Hall, N. C., & Lüdtke, O. (2007). Between- and within-domain relations of students' academic emotions. *Journal of Educational Psychology, 99*, 715–733. doi:10.1037/0022-0663.99.4.715
- Goetz, T., & Hall, N. C. (in press). Academic boredom. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *Handbook of emotions in education*. New York, NY: Taylor & Francis.
- Goetz, T., Pekrun, R., Hall, N., & Haag, L. (2006). Academic emotions from a social-cognitive perspective: Antecedents and domain specificity of students' affect in the context of Latin instruction. *British Journal of Educational Psychology, 76*, 289–308. doi:10.1348/000709905X42860
- Hall, N. C., Perry, R. P., Chipperfield, J. G., Clifton, R. A., & Haynes, T. L. (2006). Enhancing primary and secondary control in achievement settings through writing-based attributional retraining. *Journal of Social and Clinical Psychology, 25*, 361–391. doi:10.1521/jscp.2006.25.4.361
- Harackiewicz, J. M., Durik, A. M., Barron, K. E., Linnenbrink-Garcia, L., & Tauer, J. M. (2008). The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest, and performance. *Journal of Educational Psychology, 100*, 105–122.
- Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods, 16*, 319–336. doi:10.1037/a0024917



- Hidi, S., & Renninger, A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*, 111–127. doi:10.1207/s15326985ep4102\_4
- Hoffman, J. L. (2002). The impact of student cocurricular involvement on student success: Racial and religious differences. *Journal of College Student Development, 43*, 712–739.
- Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424–453. doi:10.1037/1082-989X.3.4.424
- Jarvis, S., & Seifert, T. (2002). Work avoidance as a manifestation of hostility, helplessness, and boredom. *Alberta Journal of Educational Research, 48*, 174–187.
- Larson, R. W., & Richards, M. H. (1991). Boredom in the middle school years: Blaming schools versus blaming students. *American Journal of Education, 99*, 418–443. doi:10.1086/443992
- Linnenbrink, E. A., & Pintrich, P. R. (2002). Achievement goal theory and affect: An asymmetrical bidirectional model. *Educational Psychologist, 37*, 69–78. doi:10.1207/S15326985EP3702\_2
- Linnenbrink-Garcia, L., & Pekrun, R. (2011). Students' emotions and academic engagement [Special issue]. *Contemporary Educational Psychology, 36*(1). doi:10.1016/j.cedpsych.2010.11.004
- Little, T. D., Preacher, K. J., Selig, J. P., & Card, N. A. (2007). New developments in latent variable panel analyses of longitudinal data. *International Journal of Behavioral Development, 31*, 357–365. doi:10.1177/0165025407077757
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130–149. doi:10.1037/1082-989X.1.2.130
- Mann, S., & Robinson, A. (2009). Boredom in the lecture theatre: An investigation into the contributors, moderators and outcomes of boredom amongst university students. *British Educational Research Journal, 35*, 243–258. doi:10.1080/01411920802042911
- Maroldo, G. K. (1986). Shyness, boredom, and grade point average among college students. *Psychological Reports, 59*, 395–398. doi:10.2466/pr0.1986.59.2.395
- Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective. *Perspectives on Psychological Science, 1*, 133–163. doi:10.1111/j.1745-6916.2006.00010.x
- Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*, 320–341. doi:10.1207/s15328007sem1103\_2
- Marsh, H. W., & O'Mara, A. (2008). Reciprocal effects between academic self-concept, self-esteem, achievement, and attainment over seven adolescent years: Unidimensional and multidimensional perspectives of self-concept. *Personality and Social Psychology Bulletin, 34*, 542–552. doi:10.1177/0146167207312313
- Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development, 76*, 397–416. doi:10.1111/j.1467-8624.2005.00853.x
- McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*, 577–605. doi:10.1146/annurev.psych.60.110707.163612
- Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology, 82*, 60–70. doi:10.1037/0022-0663.82.1.60
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543. doi:10.1007/BF02294825
- Mikulas, W. L., & Vodanovich, S. J. (1993). The essence of boredom. *The Psychological Record, 43*, 3–12.
- Murayama, K., Pekrun, R., Lichtenfeld, S., & vom Hofe, R. (2013). Predicting long-term growth in students' mathematics achievement: The unique contributions of motivation and cognitive strategies. *Child Development, 84*, 1475–1490. doi:10.1111/cdev.12036
- Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide*. Los Angeles, CA: Author.
- Nett, U. E., Goetz, T., & Hall, N. C. (2011). Coping with boredom in school: An experience sampling perspective. *Contemporary Educational Psychology, 36*, 49–59. doi:10.1016/j.cedpsych.2010.10.003
- Newberry, A. L., & Duncan, R. D. (2001). Roles of boredom and life goals in juvenile delinquency. *Journal of Applied Social Psychology, 31*, 527–541. doi:10.1111/j.1559-1816.2001.tb02054.x
- Pekrun, R. (1992). The expectancy-value theory of anxiety: Overview and implications. In D. G. Forgays, T. Sosnowski, & K. Wrzesniewski (Eds.), *Anxiety: Recent developments in self-appraisal, psychophysiological, and health research* (pp. 23–41). Washington, DC: Hemisphere.
- Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*, 315–341. doi:10.1007/s10648-006-9029-9
- Pekrun, R., Cusack, A., Murayama, K., Elliot, A. J., & Thomas, K. (2014). The power of anticipated feedback: Effects on students' achievement goals and achievement emotions. *Learning and Instruction, 29*, 115–124. doi:10.1016/j.learninstruc.2013.09.002
- Pekrun, R., Elliot, A. J., & Maier, M. A. (2009). Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance. *Journal of Educational Psychology, 101*, 115–135. doi:10.1037/a0013383
- Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., & Perry, R. P. (2010). Boredom in achievement settings: Control-value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology, 102*, 531–549. doi:10.1037/a0019243
- Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology, 36*, 36–48. doi:10.1016/j.cedpsych.2010.10.002
- Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of quantitative and qualitative research. *Educational Psychologist, 37*, 91–105. doi:10.1207/S15326985EP3702\_4
- Pekrun, R., & Linnenbrink-Garcia, L. (Eds.). (in press). *Handbook of emotions in education*. New York, NY: Taylor & Francis.
- Pekrun, R., & Perry, R. P. (2013). Self-processes in achievement emotions: Perspectives of the control-value theory. In D. M. McInerney, H. W. Marsh, R. Craven, & F. Guay (Eds.), *Theory driving research: New wave perspectives on self-processes and human development* (pp. 83–108). Charlotte, NC: Information Age.
- Pekrun, R., vom Hofe, R., Blum, W., Frenzel, A. C., Goetz, T., & Wartha, S. (2007). Development of mathematical competencies in adolescence: The PALMA longitudinal study. In M. Prenzel (Ed.), *Studies on the educational quality of schools* (pp. 17–37). Münster, Germany: Waxmann.
- Perry, R. P., Hladkyj, S., Pekrun, R., & Pelletier, S. (2001). Academic control and action control in college students: A longitudinal study of self-regulation. *Journal of Educational Psychology, 93*, 776–789. doi:10.1037/0022-0663.93.4.776
- Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)* (Technical Report No. 91-B-004). Ann Arbor, MI: University of Michigan.
- Rennert, H., & Berger, I. (1956). Pädagogische und kinderpsychiatrische Betrachtungen über geistig vorausentwickelte Kinder [Educational and

- psychiatric observations of children showing precocious intellectual development]. *Praxis der Kinderpsychologie und Kinderpsychiatrie*, 5, 293–296.
- Rosel, J., & Plewis, I. (2008). Longitudinal data analysis with structural equations. *Methodology*, 4, 37–50. doi:10.1027/1614-2241.4.1.37
- Roseman, W. P. (1975). Boredom at school. *British Journal of Educational Psychology*, 45, 141–152.
- Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis: A Festschrift for Heinz Neudecker* (pp. 233–247). New York, NY: Springer. doi:10.1007/978-1-4615-4603-0\_17
- Schiefele, U. (2009). Situational and individual interest. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 197–222). New York, NY: Routledge.
- Schutz, P. A., & Lanehart, S. L. (Eds.). (2002). Emotions in education [Special issue]. *Educational Psychologist*, 37(2).
- Schutz, P. A., & Pekrun, R. (Eds.). (2007). *Emotion in education*. San Diego, CA: Academic Press.
- Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464. doi:10.1214/aos/1176344136
- Sisk, D. A. (1988). The bored and disinterested gifted child: Going through school lockstep. *Journal for the Education of the Gifted*, 11, 5–18.
- Skinner, E. A., Furrer, C., Marchand, G., & Kindermann, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic? *Journal of Educational Psychology*, 100, 765–781. doi:10.1037/a0012840
- Thackray, R. I. (1981). The stress of boredom and monotony: A consideration of the evidence. *Psychosomatic Medicine*, 43, 165–176.
- Trautwein, U., Marsh, H. W., Nagengast, B., Lüdtke, O., Nagy, G., & Jonkmann, K. (2012). Probing for the multiplicative term in modern expectancy-value theory: A latent interaction modeling study. *Journal of Educational Psychology*, 104, 763–777. doi:10.1037/a0027470
- Turner, J. E., & Schallert, D. L. (2001). Expectancy–value relationships of shame reactions and shame resiliency. *Journal of Educational Psychology*, 93, 320–329. doi:10.1037/0022-0663.93.2.320
- Turner, J. E., & Waugh, R. M. (2007). A dynamical systems perspective regarding students' learning processes: Shame reactions and emergent self-organizations. In P. A. Schutz & R. Pekrun (Eds.), *Emotions in education* (pp. 125–145). San Diego, CA: Academic Press. doi:10.1016/B978-012372545-5/50009-5
- van Tilburg, W. A. P., & Igou, E. R. (2012). On boredom: Lack of challenge and meaning as distinct boredom experiences. *Motivation and Emotion*, 36, 181–194. doi:10.1007/s11031-011-9234-9
- Vogel-Walcutt, J. J., Fiorella, L., Carper, T., & Schatz, S. (2012). The definition, assessment, and mitigation of state boredom within educational settings: A comprehensive review. *Educational Psychology Review*, 24, 89–111. doi:10.1007/s10648-011-9182-7
- Watt, J. D., & Vodanovich, S. J. (1999). Boredom proneness and psychosocial development. *Journal of Psychology: Interdisciplinary and Applied*, 133, 303–314. doi:10.1080/00223989909599743
- Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, 92, 548–573. doi:10.1037/0033-295X.92.4.548
- Zeidner, M. (1998). *Test anxiety: The state of the art*. New York, NY: Plenum Press.
- Zeidner, M. (2007). Test anxiety in educational contexts: Concepts, findings, future directions. In P. A. Schutz & R. Pekrun (Eds.), *Emotions in education* (pp. 165–184). San Diego, CA: Academic Press. doi:10.1016/B978-012372545-5/50011-3
- Zheng, J. L., Saunders, K. P., Shelley, M. C., II, & Whalen, D. F. (2002). Predictors of academic success for freshmen residence hall students. *Journal of College Student Development*, 43, 267–283.

## Appendix A

### Items of the Boredom Scale

1. When studying for this course, I feel bored.
2. The things I have to do for this course are often boring.
3. The content is so boring that I often find myself daydreaming.
4. When studying, my thoughts are everywhere else, except on the course material.
5. Often I am not motivated to invest effort in this boring course.
6. The material in this subject area is so boring that it makes me exhausted even to think about it.

(Appendices continue)



## Appendix II

## Standardized Factor Loadings, Path Coefficients, and Residual Variances for Model 5

Coefficient	Boredom					Test performance				
	Time 1	Time 2	Time 3	Time 4	Time 5	Test 1	Test 2	Test 3	Test 4	Test 5
Factor loadings										
Item 1	.77**	.72**	.82**	.79**	.85**					
Item 2	.77**	.84**	.81**	.83**	.83**					
Item 3	.85**	.87**	.88**	.90**	.87**					
Item 4	.61**	.63**	.66**	.74**	.73**					
Item 5	.71**	.79**	.77**	.80**	.82**					
Item 6	.68**	.73**	.72**	.72**	.77**					
Path coefficients										
Boredom Time $n^a$						-.25**	-.09*	-.13**	-.09**	-.16**
Boredom Time $n - 1^b$		.68**	.69**	.83**	.82**					
Test $n - 1^c$		-.15**	-.11**	-.11	-.14**		.58**	.61**	.52**	.47**
Interest	-.44**	-.11*	-.15**	.10*	-.04	.07**	.03	-.01	.02	-.02
Intrinsic motivation	-.15	.02	.01	-.10*	-.01	-.13	-.02	-.06*	.01	-.08**
Gender	-.24**	-.06	-.03	-.06	.00	-.04	.01	.02	-.01	.07
Age	-.09*	.01	-.01	-.03	-.03	.03	.15**	.02	.11	.13*
High-school grades	-.01	-.01	.00	.07	.04	.37**	.20**	.18**	.15*	.25**
Residual variances	.89**	.34**	.34**	.25**	.20**	.83**	.48**	.44**	.59**	.54**

<sup>a</sup> Effects of Boredom Times 1, 2, 3, 4, and 5 on Tests 1, 2, 3, 4, and 5, respectively. <sup>b</sup> Effects of Boredom Times 1, 2, 3, and 4 on Boredom Times 2, 3, 4, and 5, respectively. <sup>c</sup> Effects of Tests 1, 2, 3, and 4 on Boredom Times 2, 3, 4, and 5 and Tests 2, 3, 4, and 5, respectively.

\*  $p < .05$ . \*\*  $p < .01$ .

Received November 30, 2012

Revision received November 14, 2013

Accepted November 25, 2013 ■

# Perfectionism and Motivation of Adolescents in Academic Contexts

Mimi Bong, Arum Hwang, Arum Noh, and Sung-il Kim  
Korea University

We examined the nature of self-oriented and socially prescribed perfectionism in relation to the motivation and achievement of 306 Korean 7th graders. We also tested the mediating role of domain-specific academic self-efficacy and achievement goals in the relationships between perfectionism and achievement-related outcomes across math and English. In the direct path model, self-oriented perfectionism related positively to academic achievement and negatively to acceptability of cheating and academic procrastination. Socially prescribed perfectionism, in contrast, related positively to test anxiety, acceptability of cheating, and academic procrastination. In the mediation models, self-oriented perfectionism related consistently and positively to academic self-efficacy, a mastery goal, and a performance-approach goal in the domain. Socially prescribed perfectionism related consistently and positively to a performance-approach and a performance-avoidance goal. Academic self-efficacy and a mastery goal mediated the paths from self-oriented perfectionism to acceptability of cheating, academic procrastination, and achievement, while a performance-avoidance goal in English mediated the path from socially prescribed perfectionism to test anxiety. Many of the paths from perfectionism to outcomes were thus mediated by domain-specific motivation. The direct paths from the 2 perfectionism dimensions to academic procrastination remained significant, however, even in the presence of the intervening motivation variables.

**Keywords:** perfectionism, self-efficacy, achievement goals, anxiety, procrastination

Perfectionism refers to the personality trait of setting difficult goals and evaluating one's own performance critically against these goals (Flett, Hewitt, & Dyck, 1989; Frost & Marten, 1990). Its strong associations with diverse symptoms of psychological maladjustment and disorders have made it the topic of extensive research in the past (Hewitt & Flett, 1991). There is reason to suspect, however, that perfectionism is a multidimensional construct and may not always be a harmful characteristic to possess (Frost, Marten, Lahart, & Rosenblate, 1990; Hewitt & Flett, 1991). For example, perfectionism influences goal-setting. Goals determine the direction of behavior, strength of effort and persistence, and quality of final performance (Locke & Latham, 2002). The motivation and achievement of perfectionists, who strive for challenging goals, would be different from those of nonperfectionists.

Despite the apparent relevance to achievement striving, only a small number of studies to date have tested how perfectionism operates in academic settings. Recent evidence demonstrates that certain forms of perfectionism could prove beneficial in learning situations (see Fletcher & Speirs Neumeister, 2012, for review).

The literature, however, is yet to offer concrete answers to questions such as how perfectionism relates to motivation in school, what the nature of relationships is between perfectionism and achievement-related outcomes, and which type of perfectionism is actually conducive to learning. The few available studies conducted in academic contexts have involved college students in North America (Mills & Blankstein, 2000; Verner-Filion & Gaudreau, 2010), which further limits generalizability of the findings to adolescent populations and other cultures.

We tried to address these issues by investigating the relationships between different types of perfectionism and indexes of academic motivation and performance in a group of Korean adolescent students. More broadly, we were interested in the extent domain-specific motivational beliefs mediated the effects of personality dispositions on achievement-related outcomes. We first examined the dimensional characteristics of perfectionism in relation to test anxiety, acceptability of cheating, academic procrastination, and achievement, to explore the nature of multidimensional perfectionism in academic contexts. We then tested the role of academic self-efficacy and achievement goals as potential mediators in these associations.

## Self-Oriented and Socially Prescribed Perfectionism

### Perfectionism as a Multidimensional Construct

Frost et al. (1990) claimed that a unidimensional definition of perfectionism, as the tendency to set excessively high personal standards, cannot distinguish highly competent and successful "normal" perfectionists from "neurotic" perfectionists. They viewed perfectionism to be multidimensional, comprising six correlated dimensions: high personal standards, a concern over mistakes in

This article was published Online First February 10, 2014.

Mimi Bong, Arum Hwang, Arum Noh, and Sung-il Kim, Department of Education and bMRI (Brain and Motivation Research Institute), Korea University, Seoul, Korea.

This research was supported by the WCU (World Class University) Program funded by the Korean Ministry of Education, Science and Technology, consigned to the Korea Science and Engineering Foundation (Grant R32-2008-000-20023-0).

Correspondence concerning this article should be addressed to Mimi Bong, Department of Education, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 136-701, Korea. E-mail: mimibong@korea.ac.kr



performance, feelings of doubt about quality of actions, valuing of parents' expectations, apprehension of parents' criticism of performance, and an overemphasis on organization, neatness, and order. All six dimensions, except for the organization dimension, correlated positively with fear of failure. While the concern over mistakes and doubts about actions dimensions correlated positively with maladaptive symptoms such as depression, obsessive-compulsive disorder, feelings of guilt, and procrastination, the personal standards and organization dimensions did not. The personal standards dimension even correlated negatively with procrastination and positively with self-efficacy. The parental expectations and parental criticism dimensions are now considered antecedents rather than functional dimensions of perfectionism (Fletcher, Shim, & Wang, 2012).

Hewitt and Flett (1991) also viewed perfectionism as a multi-dimensional construct but with a different set of dimensions. They claimed that there are three dimensions of perfectionism that interact with different types of stressors to produce distinct outcomes: self-oriented, other-oriented, and socially prescribed perfectionism. Self-oriented perfectionists impose high standards upon themselves, evaluate their own performance against these standards, and strive to perform flawlessly to meet these standards. Other-oriented perfectionists enforce high standards upon others, evaluate others' performance against those standards, and insist others perform perfectly to meet those standards. These two dimensions are differentiated from socially prescribed perfectionism based on who takes charge of setting the standards. Whereas self- and other-oriented perfectionists strive to satisfy, or demand others to satisfy, the standards that they generate, socially prescribed perfectionists strive to meet the standards that significant others, such as parents, impose on them (Stoeber, Feast, & Hayward, 2009).

Because the conceptualization of Hewitt and Flett (1991) deals with both intraindividual and interpersonal aspects of perfectionism, it appears more pertinent to the study of children and adolescents than that of Frost et al. (1990). Among the three types of perfectionism, other-oriented perfectionism seems least relevant because children and adolescents are more often targets of other-oriented perfectionism than they are other-oriented perfectionists themselves. We were thus only interested in self-oriented and socially prescribed perfectionism in this research.

### Perfectionism in East Asian Cultures

Because the present sample consisted of Korean adolescents, it is important to inspect at the outset features in East Asian cultures that may render the distinction between self-oriented and socially prescribed perfectionism particularly consequential. Collectivism is one such feature. Individuals in countries such as Korea, China, and Japan tend to embrace interdependent self-construal (Heine, 2001; Markus & Kitayama, 1991; Oishi & Diener, 2001). They strive hard to maintain group harmony by paying keen attention to in-group members' feelings, opinions, and actions, trying to please and to avoid displeasing significant others, and conforming to established norms and standards. For adolescents in East Asian cultures, judgments of success and failure in school would depend heavily on what parents, teachers, and society in general deem satisfactory.

Another relevant feature in East Asian cultures is the high standards of academic excellence that parents impose on their child. In a study by Okagaki and Frensch (1998), Asian parents reported significantly higher "expected" as well as "ideal" educational attainments for their child than did European American and Latino parents. Asian parents also displayed significantly stronger negative reactions to the hypothetical scenarios of their child receiving grades of B's and C's instead of A's. At the same time, children in East Asian cultures have a strong sense of gratitude and indebtedness to their parents (Park & Kim, 2006). A strong sense of obligation coupled with high parental standards could increase socially prescribed perfectionism in Asian students.

Castro and Rice (2003) reported that Asian American college students indeed scored significantly higher on Frost et al.'s (1990) perfectionism dimensions of parental criticism, concerns over mistakes, and doubts about actions than did European and African American students. They also scored significantly higher on parental expectations and personal standards than did European American students. The two parental dimensions are strong correlates of socially prescribed perfectionism (Flett, Sawatzky, & Hewitt, 1995). Furthermore, perfectionism accounted for 27% of the variance in Asian American students' grade-point averages (GPAs), compared to only 7% in European American students' GPAs. Socially prescribed perfectionism is hence judged to be a particularly meaningful construct to examine in relation to Asian students' school achievement.

### Perfectionism in Academic Contexts

#### Perfectionism as a Predictor of Achievement-Related Outcomes

Self-oriented and socially prescribed perfectionism typically demonstrate moderate to strong positive correlations to each other. Even so, socially prescribed perfectionism correlates with a broader array of psychological maladjustments than does self-oriented perfectionism. In Hewitt et al. (2002), for example, both types of perfectionism correlated positively with anxiety and depression. Socially prescribed perfectionism further correlated positively with outward expression of anger and social stress and negatively with anger suppression. Based on these results, Hewitt et al. concluded that both self-oriented and socially prescribed perfectionism make children and adolescents vulnerable to maladjustment, albeit to differential degrees.

A picture coming out of the academic domain, however, is somewhat different. Noting the unambiguous bearing of perfectionism on achievement behavior, several investigators have tried to unearth the psychological and behavioral profiles associated with each perfectionism dimension in academic settings. Because perfectionists' striving to attain difficult goals often results in negative affect and counterproductive behavior (Einstein, Lovibond, & Gaston, 2000), variables such as anxiety and procrastination, along with achievement, have been closely examined in relation to perfectionism. Anxiety and procrastination are major impediments to successful coping and performance (Steel, 2007; Zeidner, 1994). It is important to learn, therefore, if perfectionism actually elevates these negative psychological and behavioral responses in achievement situations and, if so, why.



Unfortunately, relationships of the two perfectionism dimensions with anxiety and procrastination have been less than straightforward. In a study by Mills and Blankstein (2000), both self-oriented and socially prescribed perfectionism correlated positively with test anxiety and extrinsic motivation, consistent with the observation of Hewitt et al. (2002). However, self-oriented perfectionism in their study also correlated positively with adaptive motivation and learning process variables such as self-efficacy, task value, use of various cognitive strategies, and effective resource management. Socially prescribed perfectionism did not correlate significantly or correlated negatively with these variables. In Einstein et al. (2000), only socially prescribed perfectionism correlated with anxiety and depression, although both perfectionism dimensions correlated positively with stress.

Procrastination is another variable frequently studied in relation to multidimensional perfectionism. Socially prescribed perfectionism correlates with a general tendency of procrastination as well as academic procrastination (Flett, Blankstein, Hewitt, & Koledin, 1992). Negative perfectionism, which is analogous to socially prescribed perfectionism, also correlates positively with academic procrastination, while positive perfectionism, which is analogous to self-oriented perfectionism, does not (Burns, Dittmann, Nguyen, & Mitchelson, 2000). Contrary to these findings, a meta-analysis by Steel (2007) showed that the perfectionism-procrastination correlation was negligible ( $r = -.03$ ). This discrepancy likely owes to the definition of perfectionism in Steel's review, which comprised only of self- and other-oriented perfectionism. Socially prescribed perfectionism was classified as an index of fear of failure, which did correlate positively with procrastination ( $r = .18$ ). Further empirical tests will help clarify the relationship between perfectionism and procrastination.

Researchers have also been interested in the relationship between perfectionism and cheating for obvious reasons. Cheating provides a means to attain an otherwise impossible goal. Earlier, we described that Asian parents put high academic demands on their child (Okagaki & Frensch, 1998), and Asian students, in turn, perceive high parental expectations and parental criticism (Castro & Rice, 2003) that are antecedents of socially prescribed perfectionism (Flett et al., 1995). High parental pressure functions as a source of conflict between Korean parents and children, which increases acceptability of cheating behavior among Korean adolescents (Bong, 2008). Still, the results on cheating have not been fully consistent, either. Vansteenkiste et al. (2010) found that the personal standards dimension of perfectionism correlated negatively with acceptability of cheating as well as actual cheating behavior. The concern over mistakes and doubts about actions dimensions correlated with neither of them. Nathanson, Paulhus, and Williams (2006), however, failed to find a significant relationship between self-oriented or socially prescribed perfectionism and cheating behavior.

When it comes to academic performance, being self-oriented perfectionists clearly helps. Bieling, Israeli, Smith, and Antony (2003) reported that adaptive perfectionism, which included self-oriented perfectionism, correlated positively with both positive and negative affect toward the recent exam, future plans to study more, grade goals for the current and future exams, and actual exam performance. Maladaptive perfectionism, which included socially prescribed perfectionism, also correlated positively with negative affect toward the exam. However, unlike adaptive perfectionism, it

correlated negatively with positive affect toward the exam or exam preparedness. Other studies similarly depict the performance benefits of self-oriented perfectionism. Stoeber and Rambow (2007) observed that students with an adaptive form of perfectionism attained significantly higher academic achievement compared to those with a maladaptive form of perfectionism. Verner-Filion and Gaudreau (2010) also reported that self-oriented perfectionism positively predicted academic satisfaction and grade point averages for college students, whereas socially prescribed perfectionism negatively predicted them.

Given the negative effect anxiety has on performance, it is puzzling that self-oriented perfectionism, which often correlates positively with anxiety, enhances performance. The answer may come from the trait-state distinction. When Zeidner (1994) examined the relationships between multiple components of trait anxiety and state anxiety, only social evaluation trait anxiety predicted state anxiety before the exam, directly and indirectly via academic stress. In Mills and Blankstein's (2000) study described earlier, self-oriented perfectionism no longer correlated with test anxiety, when its covariance with socially prescribed perfectionism was controlled for. These results suggest that socially prescribed perfectionism increases state anxiety, while self-oriented perfectionism does not, even though both correlate with trait anxiety (Flett, Hewitt, Endler, & Tassone, 1994). This conjecture requires a mediating process that weakens the link of self-oriented perfectionism to state anxiety, which we describe in the next section.

To summarize, self-oriented perfectionism demonstrates positive associations with academic achievement and null or negative associations with test anxiety, academic procrastination, and acceptability of cheating. Socially prescribed perfectionism, on the contrary, demonstrates negative associations with achievement and positive associations with detrimental indexes in academic settings.

### Academic Motivation as a Mediator Between Perfectionism and Outcomes

Self-oriented perfectionism, therefore, appears to play at least a more positive than negative function in the learning process. However, more evidence is needed to conclude that it is indeed an adaptive form of perfectionism for learners in academic contexts. In addition, most of the few available studies simply contrasted relationships of the two perfectionism dimensions with various outcomes without probing why they were associated with different outcomes or with the same outcomes in different manners.

Miquelon, Vallerand, Grouzet, and Cardinal (2005) argued that failure to integrate mediating motivational processes in the relationships between perfectionism and outcomes has been responsible for the ambiguous effects associated with self-oriented perfectionism. In their study, self-oriented and socially prescribed perfectionism for college students correlated positively with each other as well as with neuroticism (Study 2). When motivational constructs were incorporated as mediators in path analysis, however, the two displayed completely different predictive patterns. Self-oriented perfectionism positively predicted self-determined academic motivation, which in turn positively predicted academic adjustment and negatively predicted psychological adjustment difficulties. Socially prescribed perfectionism, on the contrary, pos-



itively predicted non-self-determined academic motivation, which in turn positively predicted psychological adjustment difficulties.

Seo (2008) also tested the mediating role of academic self-efficacy in the relationship of self-oriented perfectionism with academic procrastination. Self-oriented perfectionism correlated positively with self-efficacy and negatively with procrastination. More important, academic self-efficacy fully mediated the relationship between self-oriented perfectionism and academic procrastination. These studies corroborate the adaptive nature of self-oriented perfectionism and illustrate the mediating role academic motivation plays in the perfectionism-outcome links.

## Perfectionism and Academic Motivation

### Perfectionism and Academic Self-Efficacy

Among many motivational constructs, one plausible mediator between perfectionism and learning outcomes is academic self-efficacy. Self-efficacy represents subjective convictions for successfully carrying out courses of action to achieve desired outcomes (Bandura, 1977). Beliefs of self-efficacy are tailored to particular tasks, activities, or domains of functioning. Academic self-efficacy, therefore, refers to learners' subjective convictions for successfully performing specific academic tasks at designated levels (Schunk, 1991). The central role academic self-efficacy plays in determining the strength of motivation and quality of achievement-related outcomes in so many different settings and subject areas (Multon, Brown, & Lent, 1991; Pajares, 1996) strongly suggests that self-efficacy functions as a mediator between stable personality characteristics such as perfectionism and outcomes in specific learning contexts.

Perfectionism and academic self-efficacy would most likely be intertwined with each other through the psychological mechanisms of goal-setting and self-evaluation. Self-oriented perfectionism, by definition, involves setting high goals and striving to attain them (Hewitt & Flett, 1991). Bandura (1997) asserted that acts of setting and pursuing challenging personal goals and aspirations foster development of self-efficacy, a claim that has received strong empirical support from both self-efficacy and goal-setting literatures. As individuals pursue higher goals, their self-efficacy and performance improve correspondingly (Locke & Latham, 2002). Academic self-efficacy mediates the connection between goals and eventual performance as students work toward their goals, monitoring their progress and developing necessary skills (Schunk, 1996). Self-efficacy of learners is best promoted when they set challenging goals and engage in frequent self-evaluations of their goal progress (Schunk & Ertmer, 1999).

Self-oriented perfectionists, who are in pursuit of difficult self-set goals, would be vigilant about assessing their performance against these goals because goals also serve as standards with which to evaluate performance (Locke & Latham, 2002). Accomplishment of proximal subgoals while striving to achieve the difficult final goal provides these perfectionistic learners with mastery experiences, which constitute the most potent source of self-efficacy information (Bandura, 1977). The end product is a stronger sense of self-efficacy, accompanied by intrinsic interest, self-satisfaction, and enhanced performance (Bandura & Schunk, 1981).

Trying to satisfy difficult standards is a hallmark of not only self-oriented perfectionism but also socially prescribed perfectionism. Yet socially prescribed perfectionists do not necessarily enjoy the profit of goal pursuit in the form of improved self-efficacy. According to Latham and Locke (1991), the effects of goal-setting are moderated by the degree of goal commitment. For individuals who are high in goal commitment, performance improves linearly with goal difficulty, presumably with the help of augmented perceptions of self-efficacy. For those who are low in goal commitment, however, performance shows no systematic relationship with goal difficulty. Goal commitment is higher when the goals are attainable and individuals participate in setting them, compared to when they are impossible to attain and assigned by others. Socially prescribed perfectionists strive to fulfill excessively high standards that are imposed by others. Goal commitment of socially prescribed perfectionists, therefore, would not be as strong as that of self-oriented perfectionists and the weaker goal commitment compromises the self-efficacy benefit they should otherwise reap from their enactive mastery experiences.

Supporting these conjectures, Mills and Blankstein (2000) observed that self-oriented perfectionism demonstrated a positive correlation with academic self-efficacy for learning and performance in an introductory psychology course. Socially prescribed perfectionism exhibited a nonsignificant correlation with academic self-efficacy, which became significant and negative when only the unique variance was considered. Similarly, Van Yperen (2006) reported that a subdimension of self-oriented perfectionism, the importance of being perfect, correlated positively with self-efficacy, while a subdimension of socially prescribed perfectionism, others' high standards, correlated negatively with it.

Evidence of mediation by self-efficacy was also observed in a study by Dunkley, Zuroff, and Blankstein (2003). The researchers hypothesized that self-blame, lower self-efficacy, and perceived criticism from others would mediate the relationship between self-critical perfectionism and avoidant coping. Personal standards perfectionism, an adaptive form of perfectionism that included self-oriented perfectionism, was distinguished from self-critical perfectionism, a maladaptive form of perfectionism that included socially prescribed perfectionism. Only self-critical perfectionism displayed a significant negative correlation with self-efficacy. Supporting the authors' hypothesis, higher self-critical perfectionism predicted lower self-efficacy, which in turn predicted greater avoidant coping in the form of denial and disengagement from stressful events. As the latter two studies assessed self-efficacy for life events, a direct test of academic self-efficacy as a mediator between perfectionism and learning outcomes is required.

### Perfectionism and Achievement Goals

Achievement goals have received even greater attention than academic self-efficacy has as potential mediators of perfectionism-outcome relationships in the academic domain. Achievement goals represent underlying purposes of achievement-related behavior in specific achievement situations (Dweck & Leggett, 1988). Although disagreement exists on the exact definition and functions of each achievement goal, a mastery goal has emerged as a positive predictor of interest, while a performance-approach goal of pursuing normative competence has emerged as a positive predictor of performance. A performance-avoidance goal of avoiding norma-



tive incompetence has been a consistent negative predictor of both outcomes (Hulleman, Schrager, Bodmann, & Harackiewicz, 2010).

A number of parallels between the literatures on perfectionism and achievement goals delineate how these constructs may be relevant to each other. Self-oriented perfectionism arises from a motive to achieve, while socially prescribed perfectionism results from fear of failure (Speirs Neumeister, 2004). An achievement motive is also an antecedent of mastery and performance-approach goals, while fear of failure is an antecedent of performance-approach and performance-avoidance goals (Elliot & Church, 1997). Sharing the same motive, self-oriented perfectionists would more likely pursue the two approach-oriented goals, whereas socially prescribed perfectionists would more likely pursue the two performance-oriented goals.

Speirs Neumeister (2004) interviewed gifted college students identified as either self-oriented or socially prescribed perfectionists and found support for the hypothesized links. Self-oriented perfectionists were driven by a strong achievement motive and adopted either a mastery goal of learning new things and improving oneself or a performance-approach goal of doing better than others. These students sought out challenging academic tasks and prepared far in advance for assignments and exams. Socially prescribed perfectionists, in comparison, were driven by a strong fear of failure and adopted either a performance-approach goal of validating one's ability or a performance-avoidance goal of avoiding doing worse than others. These students procrastinated to exonerate themselves from the implications of potential failure.

The two perfectionism dimensions exhibited different associations with measures of self-criticism as well (Trumpeter, Watson, & O'Leary, 2006). On the one hand, both perfectionism dimensions correlated positively with internalized self-criticism, a negative evaluation of the self due to failure to meet self-set standards. On the other hand, only socially prescribed perfectionism correlated positively with comparative self-criticism, a negative evaluation of the self due to failure to perform as well as others. These findings suggest that the nature of competence evaluation associated with each perfectionism dimension might provide another mechanism through which perfectionism promotes particular achievement goals. Self-oriented perfectionists, who evaluate their performance solely against self-set standards, would likely be drawn to a mastery goal, which defines competence in an absolute sense in reference to goals and standards (Elliot & McGregor, 2001; Pintrich, 2000). Socially prescribed perfectionists, who evaluate their performance against that of others, would find performance-approach and performance-avoidance goals more attractive because the normative definition of competence (Elliot & McGregor, 2001; Pintrich, 2000) aligns well with the way socially prescribed perfectionists evaluate competence.

Verner-Filion and Gaudreau (2010) not only replicated the proposed links between perfectionism and achievement goals for college students but also presented evidence of mediation by achievement goals. They assessed perfectionism and achievement goals before midterm exams and academic satisfaction and grade point averages after midterm exams. Self-oriented perfectionism positively linked to mastery, performance-approach, and performance-avoidance goals. It also positively predicted academic satisfaction and grade point averages. Socially prescribed perfectionism negatively linked to mastery goals and positively

linked to performance-approach and performance-avoidance goals. It negatively predicted academic satisfaction and grade point averages. The paths between perfectionism and academic satisfaction were mediated by a mastery goal and those between perfectionism and grade point averages were mediated by a performance-approach goal.

More specifically, as students' self-oriented perfectionism became stronger, they were better positioned to experience improved academic satisfaction and higher grade point averages. Adoption of a mastery goal provided one channel through which self-oriented perfectionism resulted in increased academic satisfaction, while adoption of a performance-approach goal resulted in higher academic achievement for self-oriented perfectionism. Quite the contrary, as students' socially prescribed perfectionism became stronger, they experienced decreased academic satisfaction and lower academic achievement. Socially prescribed perfectionism made pursuit of a mastery goal less likely, which partly explained why it was associated with reduced academic satisfaction. However, socially prescribed perfectionism also meant a greater likelihood of adopting a performance-approach goal, which predicted higher, not lower, subsequent grade point averages. These results illustrate complex routes by which different types of perfectionism connect to achievement-related outcomes and highlight the benefits of incorporating motivational variables such as achievement goals in the relationships between perfectionism and learning outcomes.

### **Self-Efficacy and Achievement Goals as Predictors of Achievement-Related Outcomes**

In our review of the literature presented earlier, we focused on test anxiety, academic procrastination, cheating, and academic performance among diverse outcomes that perfectionism predicts in learning situations, due to their direct implications for students' achievement striving. These achievement-related outcomes are also the ones that self-oriented and socially prescribed perfectionism have demonstrated contrasting associations in past research. Evidence diverges on several of these relationships, however, which would benefit from additional empirical tests. The relationships that self-efficacy and each of the achievement goals display with the same outcomes, in comparison, have been far more consistent.

According to the extant literature, academic self-efficacy is a negative predictor of test anxiety (Bandalos, Finney, & Geske, 2003) and academic procrastination (Steel, 2007; Wolters, 2003, 2004) and a positive predictor of achievement (Bong, 2005; Wolters, 2003, 2004). A mastery goal is a negative predictor of test anxiety (Bandalos et al., 2003), acceptability of cheating (Murdock, Miller, & Kohlhardt, 2004), and academic procrastination (Wolters, 2004), while a performance-approach goal is a positive predictor of acceptability of cheating (Murdock et al., 2004) and achievement (Daniels et al., 2009; Hulleman et al., 2010; Wolters, 2004). In Murdock et al. (2004), performance-approach and performance-avoidance goals formed a single factor. It is possible, therefore, that the avoidance component was primarily responsible for the positive path from the performance goal to acceptability of cheating in their study. However, Anderman, Griesinger, and Westerfield (1998) showed that an extrinsic goal, a variant of a



performance-approach goal, was also a significant positive predictor of acceptability of cheating.

The link between a performance-approach goal and anxiety is mixed, with early studies that did not distinguish between approach and avoidance components reporting a positive path (Bandalos et al., 2003; Daniels et al., 2009) and later studies reporting a nonsignificant path (Sideridis, 2005). A performance-avoidance goal is a positive predictor of anxiety (Pekrun, Elliot, & Maier, 2006; Sideridis, 2005), acceptability of cheating (Bong, 2008; Murdock et al., 2004), and academic procrastination (Wolters, 2004) and a negative predictor of achievement (Bong, 2005; Hulleman et al., 2010; Sideridis, 2005).

Regarding the relationships among the motivational constructs, different opinions exist in the literature regarding the causal precedence between academic self-efficacy and achievement goals. Dweck and Leggett (1988) viewed perceived competence mainly as a moderator of the achievement goal effects. Others treat self-efficacy as an outcome of achievement goals (e.g., Middleton & Midgley, 1997). Self-efficacy theorists maintain that a core component of self-efficacy is perceived competence (Bong & Skaalvik, 2003; Schunk & Pajares, 2005), which achievement goal theorists recognize as an antecedent of all achievement goals (e.g., Elliot, & Church, 1997). From a theoretical standpoint, then, it is most plausible to regard self-efficacy as causally predominant to achievement goals. Supporting this conjecture, changes in the academic self-efficacy of Korean high school students predicted changes in their subsequent achievement goals but not the other way around (Bong, 2005). Further, self-efficacy in the preceding semester was a significant predictor of the mastery and performance-avoidance goals in the following semester but not of the performance-approach goal. Based on these findings, we expected academic self-efficacy to precede achievement goals in our model.

### Present Hypotheses

We tried to address two primary research questions in this study: (a) Is self-oriented perfectionism adaptive, and socially prescribed

perfectionism maladaptive, in the academic domain? and (b) Do achievement goals and academic self-efficacy mediate the relationships between perfectionism and achievement-related outcomes? Assuming that the answer to Question b is yes, we were also interested in uncovering the nature of the mediation by each motivational construct in the perfectionism-outcome associations.

To answer Question a, which would help ascertain the dimensional characteristics of perfectionism, we tested the direct relationships between perfectionism and outcome variables. The left panel of Figure 1 presents the hypothesized paths. The paths that have consistent support in the literature or were expected on theoretical grounds are indicated with solid lines, whereas those that lack consistent support, with the possibility of a null relationship, are indicated with dotted lines. To answer Question b, which would substantiate the hypothesized mediation of the perfectionism effects by motivational constructs and clarify the nature of such mediation, we tested the indirect relationships between perfectionism and outcome variables via academic self-efficacy and achievement goals. The right panel of Figure 1 presents the hypothesized mediation. Again, consistency of theoretical and empirical support in the literature for each hypothesized path is indicated by solid and dotted lines.

Our main interest in this research was in the mediation model. If perfectionism maintains its direct links to the various achievement-related outcomes in the model, it may mean that the effects of stable personality characteristics on academic outcomes are too strong to be mediated by domain-specific motivation. Although such results would be inconsistent with the contemporary literature on motivation (e.g., Elliot & Church, 1997), perfectionism might just be one exception to this trend. Conversely, if domain-specific motivational constructs successfully mediate the paths between perfectionism and outcomes, this would once again highlight the functional centrality of motivational beliefs in determining the learning outcomes in specific achievement situations (e.g., Pajares, 1996).

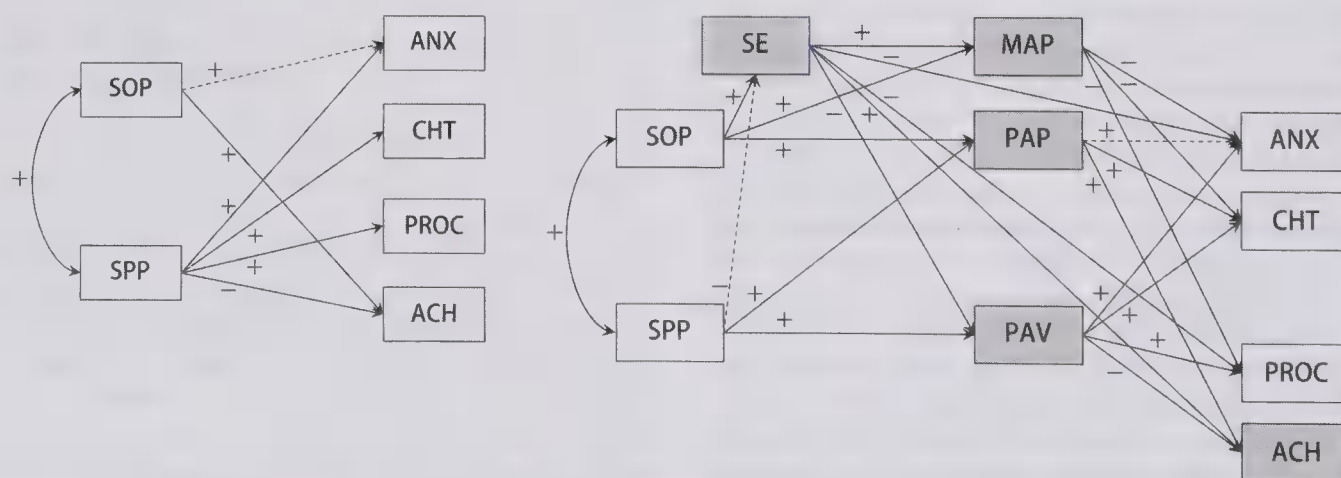


Figure 1. Hypothesized positive (+) and negative (-) paths and mediation by academic self-efficacy and achievement goals. Dotted paths indicate a possibility of nonsignificant relationships. Shaded boxes indicate variables assessed in reference to math and English. SOP = self-oriented perfectionism; SPP = socially prescribed perfectionism; SE = academic self-efficacy; MAP = mastery goal; PAP = performance-approach goal; PAV = performance-avoidance goal; CHT = acceptability of cheating; ANX = test anxiety; PROC = academic procrastination; ACH = achievement scores.

By testing complex interrelationships between perfectionism, motivation, and major achievement-related outcomes, we were also hoping that the results might shed light on the nature of not only perfectionism but also a performance-approach goal. Currently, a performance-approach goal is associated with mixed effects but the reasons behind its positive and negative effects have not been clearly understood (Senko & Harackiewicz, 2005). The way with which each perfectionism dimension predisposes students to adopt a particular achievement goal in academic settings could allow us to generate inferences regarding one such mechanism.

## Method

### Participants and Procedures

Data were collected from 306 seventh graders attending a public middle school in a metropolitan city near Seoul, Korea. This school serves middle-income families and is large in scale with 10 to 12 classes at each grade-level. Ages of the participants ranged from 12 years and 5 months to 13 years and 4 months at the time of the survey. Education in 6 years of elementary school and 3 years of middle school is compulsory in Korea. The seventh grade marks the first year after the transition to middle school.

Middle school students take a nationwide, standardized competency test at the end of their senior year, which they must pass to advance to academic-track high schools. Scores on this test also determine their eligibility to enter select high schools. Flett et al. (1994) showed that relationships of self-oriented and socially prescribed perfectionism with other variables change depending on the degree of perceived evaluative threat. Because we were interested in the function of perfectionism under normal academic circumstances, seventh graders who were yet to experience elevated test stress were deemed an appropriate target for this research.

Surveys were administered during regular classroom hours, several days before final exams. We assured students of confidentiality of their responses. Data from 304 students (148 girls, 156

boys) were analyzed, excluding two students with too many missing responses.

### Measures

Students responded to items on a 5-point Likert-type scale with 1 indicating *strongly disagree* and 5 *strongly agree* for all but the academic procrastination scale, which had 1 indicating *never procrastinate* and 5 *always procrastinate*. All scales had been translated and validated in Korean in previous research (see below). The Cronbach's  $\alpha$ s obtained in the present study are reported in Table 1.

Academic self-efficacy and achievement goals were assessed in reference to the specific academic subjects of math and English because, (a) whereas academic motivation is generally domain-specific, academic self-efficacy (Bong, 1997; Pajares, 1996) and achievement goals (Bong, 2001) contain particularly strong domain-specific components; (b) motivation in math and that in English are distinct from each other (Bong, 1997; Marsh, Byrne, & Shavelson, 1988); and (c) testing the hypothesized mediation across two discrete subject matter areas would help ascertain generalizability of the hypothesized mediation.

**Self-oriented and socially prescribed perfectionism.** We used the Multidimensional Perfectionism Scale (MPS) by Hewitt and Flett (1991). The scale contains 15 items assessing self-oriented perfectionism (e.g., "I demand nothing less than perfection of myself") and another 15 assessing socially prescribed perfectionism (e.g., "The people around me expect me to succeed at everything I do"). Both scales had demonstrated satisfactory internal consistency with  $\alpha$ s above .85 in past research (Flett et al., 1995; Miquelon et al., 2005; Stoeber et al., 2009). The translated versions had functioned well among Korean college students as well (Seo & Synn, 2006;  $\alpha$ s = .88 for self-oriented perfectionism and .77 for socially prescribed perfectionism).

**Academic self-efficacy.** We used five items in the self-efficacy subscale of the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich & De Groot, 1990) to assess academic self-efficacy (e.g., "I'm certain I can understand the ideas taught in

Table 1  
*Descriptive Statistics and Reliability of Scales*

Variable	<i>M</i>	<i>SD</i>	$\alpha$	Range	Min. observed	Max. observed	Skewness	Kurtosis
Self-oriented perfectionism	3.44	0.60	.82	1-5	1.67	4.93	0.05	-0.27
Socially prescribed perfectionism	3.25	0.64	.73	1-5	1.77	4.90	0.07	-0.52
Test anxiety	3.59	1.03	.79	1-5	1.00	5.00	-0.40	-0.58
Acceptability of cheating	1.81	0.89	.67	1-5	1.00	5.00	1.09	0.67
Academic procrastination	2.95	0.78	.63	1-5	1.00	5.00	0.20	-0.10
Academic self-efficacy in math	3.12	1.17	.94	1-5	1.00	5.00	0.01	-1.00
Academic self-efficacy in English	3.45	1.10	.92	1-5	1.00	5.00	-0.40	-0.70
Mastery goal in math	4.00	0.95	.80	1-5	1.00	5.00	-1.05	1.06
Mastery goal in English	4.02	0.85	.75	1-5	1.00	5.00	-0.94	1.02
Performance-approach goal in math	3.25	1.22	.90	1-5	1.00	5.00	-0.23	-0.81
Performance-approach goal in English	3.23	1.12	.86	1-5	1.00	5.00	-0.28	-0.60
Performance-avoidance goal in math	2.69	1.18	.80	1-5	1.00	5.00	0.12	-0.94
Performance-avoidance goal in English	2.59	1.10	.76	1-5	1.00	5.00	0.23	-0.68
Achievement score in math	57.44	27.21		0-100	3.80	100.00	-0.10	-1.27
Achievement score in English	69.83	24.48		0-100	12.00	100.00	-0.74	-0.55

Note. Min. = minimum; Max. = maximum. *N* = 304.



my *subject* class"). The translated version had demonstrated  $\alpha$ s above .88 in various samples of Korean middle and high school students and correlated significantly with achievement goals, test anxiety, strategy use, perceived classroom goal structures, and academic achievement (Bong, 2005, 2008, 2009).

**Achievement goals.** Nine items were adopted from the achievement goal scale used in Elliot and McGregor (2001), which has three items each for mastery (e.g., "I want to learn as much as possible from this *subject* class"), performance-approach (e.g., "It is important for me to do better than other students in this *subject* class"), and performance-avoidance goals (e.g., "I just want to avoid doing poorly in this *subject* class"). The achievement goal scale had displayed  $\alpha$ s between .77 and .91 for the mastery goal, .90 and .94 for the performance-approach goal, and .73 and .89 for the performance-avoidance goal in their study. Scores on the scales had correlated significantly with motive dispositions, implicit theories of ability, perceived competence, anxiety, study strategies, and academic performance in past research (Cury, Elliot, Da Fonseca, & Moller, 2006; Elliot & McGregor, 2001; Pekrun et al., 2006). The translated scales had shown  $\alpha$ s ranging from .74 to .84, .61 to .92, and .65 to .78 for the mastery, performance-approach, and performance-avoidance goals, respectively, among Korean adolescents in various school subjects (Bong, 2005, 2008, 2009).

**Test anxiety.** We used the six-item test anxiety subscale of the MSLQ (Duncan & McKeachie, 2005; e.g., "When I take a test, I think about how poorly I am doing compared with other students"). The scale had proven internally consistent with  $\alpha = .80$  among college students (Mills & Blankstein, 2000; Pintrich, Smith, Garcia, & McKeachie, 1993). In previous research with sixth graders in the United States (Middleton & Midgley, 1997), the internal consistency estimate of this scale dropped to .68. The translated version had demonstrated a similar degree of internal consistency among Korean middle school students with  $\alpha = .63$  (Bong, 2009).

**Acceptability of cheating.** Three items in a scale used by Anderman et al. (1998) were modified to investigate cheating on tests (e.g., "Is it okay to cheat on tests?") because cheating is most frequently discussed in test-taking contexts in Korea. The scale had displayed acceptable internal consistency with  $\alpha$ s above .64 in U.S. middle and high school samples (Anderman et al., 1998; Murdock et al., 2004). The translated version had shown similar internal consistency estimates with  $\alpha = .64$  among Korean middle school students in past research (Hwang, 2010).

**Academic procrastination.** We administered the Procrastination Assessment Scale—Student (PASS; Solomon & Rothblum, 1984) that assesses the frequency (e.g., "Studying for exams: To what degree do you procrastinate on this task?") and perceived severity of academic procrastination (e.g., "To what degree is procrastination on this task a problem for you?"). We used only the frequency items in this study because the severity items behaved differently ( $\lambda$ s  $\leq .20$ ). The scale had exhibited internal consistency estimates above .75 and correlated significantly with perfectionism, anxiety, and grade point averages in past research (Fritzsche, Young, & Hickson, 2003; Howell, Watson, Powell, & Buro, 2006; Milgram, Marshevsky, Sadeh, 1995). The translated version had demonstrated  $\alpha = .87$  among Korean college students (Synn, Park, & Seo, 2005). We only used tasks that are applicable to middle school contexts and modified the descriptions to make them more meaningful to middle school students. These tasks included com-

pleting homework (revised from "writing a term paper"), studying for exams, and keeping up weekly class materials (revised from "keeping up weekly reading assignments").

**Academic achievement.** Final exam scores in math and English served as indexes of academic achievement. Scores on these exams could range from 0 to 100.

## Overview of Analysis

Because our sample size was considered small for the number of parameters to be estimated in the model, we applied a three-stage approach for reducing the participants-to-parameters ratio to an acceptable level, so as not to obtain nonconvergent or improper solutions (Anderson & Gerbing, 1984). We first performed preliminary confirmatory factor analyses (CFAs) per construct. For evaluating model fit, we used several fit indexes in addition to the chi-square statistics, which are known to be sensitive to sample size. We applied the Tucker–Lewis index (TLI) greater than .90 (Bentler, 1990; Tucker & Lewis, 1973), the comparative fit index (CFI) greater than .95 (Hu & Bentler, 1999), and the root-mean-square error of approximation (RMSEA) less than .08 (Browne & Cudeck, 1993) as cutoff criteria for acceptable model fit.

Using the factor loadings, factor variances, and error variances and covariances from these models, we computed factor rho coefficients as reliability estimates (Raykov, 1997, 2004). We then created a reliability-driven composite score for each latent variable (Bentler, 2009) by fixing the error variance with a formula,  $(1 - \text{scale reliability}) \times \text{scale variance}$  (Hayduk, 1987). Using these composite scores, corrected for unreliability, in subsequent analyses substantially reduced the number of parameters to be estimated to a level appropriate to our sample size.

To answer Questions a and b presented above, we first ran a direct path model between perfectionism and outcomes. When the direct paths from perfectionism to outcome variables were significant, we proceeded to test a mediating model in which the paths from perfectionism to outcome variables were mediated by academic self-efficacy and achievement goals, as illustrated in Figure 1. Because academic self-efficacy and achievement goals were assessed separately in math and English, we tested two models, one with math-specific variables and the other with English-specific ones. The statistical significance of total indirect effects, involving all mediation paths linking one variable to the other, was tested by a bootstrapping method with 1,000 bootstrapping samples with 95% bias-corrected confidence intervals. When the total indirect effects proved significant, Sobel tests followed to examine the statistical significance of individual indirect paths involved (Kline, 2005, p. 162). All measurement models and path analyses were run with AMOS 7.0 (Arbuckle, 2006).

## Results

### Descriptive Statistics

Responses to negatively worded items were reverse-coded so that high scores represent greater possession of the construct under investigation. Skewness and kurtosis statistics indicated that responses to all items approximate normal distributions. Frequency of missing responses per item ranged between 0 and 4, with missing rates less than 1.3% across all items. Missing values were

imputed with series means. Table 1 reports descriptive statistics of the scales.

Mean scores of most scales ranged between 3 and 4 on a 1–5 response scale with no strong hint of floor or ceiling effects. The acceptability of cheating scale was an exception to this trend with  $M = 1.81$ . Given the socially undesirable nature of this variable, it is not surprising that students provided low agreement ratings on these items. Responses to all scales showed acceptable degrees of internal consistency as presented in Table 1, except for the acceptability of cheating and academic procrastination scales, which were associated with somewhat low  $\alpha$ s. We believe the small number of items ( $n = 3$ ) on these scales was responsible for the low reliability.

## Measurement Models

**Measurement models per construct.** We performed CFA for each construct with individual items as indicators at this stage. When the number of indicators rendered the model just-identified and hence not testable by itself, we combined theoretically related constructs (e.g., the three achievement goals) in a single model to gain degrees of freedom. Error covariances were allowed between items for the same construct when both of the following conditions were met: (a) The content or wording of the respective items justified the covariance, and (b) the modification indexes suggested not only statistically significant but also substantial improvement in model fit. Error covariances were added to the model one at a time. Complete results from these preliminary CFAs are available from the first author upon request.

The model for self-oriented perfectionism demonstrated acceptable fit to the empirical data,  $\chi^2(83, N = 304) = 129.863, p < .01$  (TLI = .939, CFI = .952, RMSEA = .043). The socially prescribed perfectionism scale had five items with low factor loadings ( $\lambda$ s  $\leq .20$ ) that failed to reach statistical significance ( $p$ s  $> .05$ ). Four of them were reverse-coded items (e.g., “Those around me readily accept that I can make mistakes too”) and one was phrased in a negative way (i.e., “I find it difficult to meet others’ expectations of me”). These results indicate that the middle school respondents found these items unclear and different from the rest of the items. We thus excluded these items from further analyses. The final model fit the data well,  $\chi^2(30, N = 304) = 47.963, p < .05$  (TLI = .951, CFI = .967, RMSEA = .044).

The model for test anxiety demonstrated acceptable fit, although the RMSEA value was slightly over the cutoff criteria,  $\chi^2(4, N = 304) = 12.691, p < .05$  (TLI = .949, CFI = .980, RMSEA = .085). The model with acceptability of cheating and academic procrastination also displayed satisfactory fit indexes,  $\chi^2(8, N = 304) = 22.148, p < .01$  (TLI = .914, CFI = .954, RMSEA = .076). The model for academic self-efficacy in math fit the data reasonably well, again with the RMSEA value slightly over the cutoff criteria,  $\chi^2(4, N = 304) = 12.586, p < .05$  (TLI = .958, CFI = .979, RMSEA = .084). The model for academic self-efficacy in English demonstrated excellent fit,  $\chi^2(5, N = 304) = 8.785, p = .118$  (TLI = .993, CFI = .997, RMSEA = .050). The achievement goal model in both math,  $\chi^2(24, N = 304) = 64.372, p < .001$  (TLI = .956, CFI = .971, RMSEA = .075), and English,  $\chi^2(24, N = 304) = 47.517, p < .01$  (TLI = .967, CFI = .978, RMSEA = .057), produced satisfactory fit indexes.

**Full measurement models.** Next, we tested measurement models in math and English with all variables. Fit indexes are not computed, as these models are just-identified with only a single indicator for each latent variable. Table 2 presents correlation coefficients among the variables. In previous research, the adaptive and maladaptive characteristics of perfectionism and any mediating process were best demonstrated in either partial correlations or path analysis. Therefore, we only describe some of the notable findings here.

Consistent with prior reports, self-oriented perfectionism and socially prescribed perfectionism correlated positively with each other ( $r = .56$ ). Both types of perfectionism also correlated positively with test anxiety, but the correlation was stronger with socially prescribed perfectionism ( $r = .45$ ) than with self-oriented perfectionism ( $r = .25$ ). Neither type of perfectionism correlated with acceptability of cheating. Whereas self-oriented perfectionism correlated negatively with academic procrastination ( $r = -.37$ ) and positively with achievement in both math ( $r = .24$ ) and English ( $r = .22$ ), socially prescribed perfectionism did not correlate with academic procrastination and only correlated positively with achievement in math ( $r = .15$ ).

The two perfectionism variables also exhibited different patterns of correlations with subject-specific motivation variables. The correlations were largely consistent with the extant literature. Self-oriented perfectionism correlated positively with academic

Table 2  
Correlation Coefficients Among Variables

Variable	1	2	3	4	5	6	7	8	9	10
1. Self-oriented perfectionism	1.00	.56***	.25***	-.15	-.37***	.38***	.44***	.50***	.24**	.24***
2. Socially prescribed perfectionism	—	1.00	.45***	.01	.03	.11	.22**	.52***	.39***	.15*
3. Test anxiety	—	—	1.00	.06	-.01	-.04	.16*	.42***	.37***	.01
4. Acceptability of cheating	—	—	—	1.00	.36***	-.20**	-.26***	.04	.11	-.19**
5. Academic procrastination	—	—	—	—	1.00	-.56***	-.31***	-.17*	.09	-.34***
6. Academic self-efficacy	.35***	.12	-.09	-.18*	-.33***	1.00	.37***	.12	-.20**	.62***
7. Mastery goals	.48***	.24**	.25***	-.32***	-.42***	.45***	1.00	.43***	.13	.24***
8. Performance-approach goals	.50***	.60***	.50***	.06	-.15	.21**	.38***	1.00	.59***	.09
9. Performance-avoidance goals	.29***	.49***	.53***	.12	.02	-.21*	.09	.62***	1.00	-.10
10. Academic achievement	.22***	.11	-.01	-.18*	-.28***	.53***	.31***	.21***	-.08	1.00

Note.  $N = 304$ . Coefficients from the math model are presented above the diagonal; those from the English model below the diagonal. Dashes indicate that coefficients are presented in the upper diagonal.

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .



self-efficacy in both domains ( $r_s = .38$  in math and  $.35$  in English), while socially prescribed perfectionism did not. Both perfectionism variables demonstrated positive correlations with all three achievement goals. Nonetheless, a mastery goal correlated more strongly with self-oriented than socially prescribed perfectionism ( $r_s = .44$  vs.  $.22$  in math and  $.48$  vs.  $.24$  in English) and a performance-avoidance goal correlated more strongly with socially prescribed than self-oriented perfectionism ( $r_s = .39$  vs.  $.24$  in math and  $.49$  vs.  $.29$  in English). A performance-approach goal exhibited positive correlations with both self-oriented ( $r_s = .50$  in both math and English) and socially prescribed ( $r_s = .52$  in math and  $.60$  in English) perfectionism. With few exceptions, the overall pattern was highly similar across math and English.

### Path Model With Perfectionism and Outcomes Only

We performed path analysis with the same set of reliability-driven composite scores. Before testing the full path model with subject-specific academic self-efficacy and achievement goals as mediators, we examined only the direct paths from perfectionism to outcome variables. This model fit the data well,  $\chi^2(10, N = 304) = 28.376, p < .01$  (TLI = .908, CFI = .956, RMSEA = .078). Figure 2 presents statistically significant paths at  $p < .05$  from this model.

When the two perfectionism variables entered the regression equation together, the contrasting characteristics between them became clearer. Self-oriented perfectionism did not relate to test anxiety but related positively to academic achievement ( $\beta = .37$ ), supporting our hypothesis. Although not anticipated a priori, it also related negatively to acceptability of cheating ( $\beta = -.34$ ) and academic procrastination ( $\beta = -.66$ ). Socially prescribed perfectionism, in contrast, related positively to test anxiety ( $\beta = .45$ ), acceptability of cheating ( $\beta = .25$ ), and academic procrastination ( $\beta = .43$ ). Our hypothesis that socially prescribed perfectionism would be a positive predictor of maladaptive variables thus received support. Socially prescribed perfectionism did not relate significantly to achievement, however.

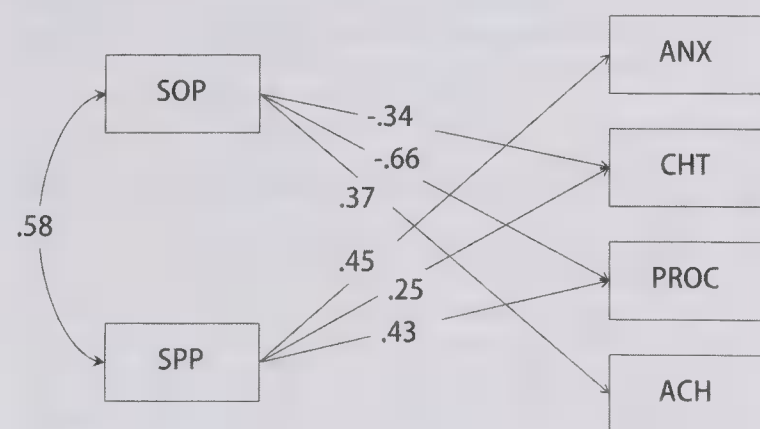


Figure 2. Path model with direct paths from perfectionism to outcomes only. Only statistically significant paths at  $p < .05$  are presented. Error terms are omitted for clarity. SOP = self-oriented perfectionism; SPP = socially prescribed perfectionism; ANX = test anxiety; CHT = acceptability of cheating; PROC = academic procrastination; ACH = achievement scores.

### Path Models With Academic Self-Efficacy and Achievement Goals as Mediators

Next, we tested full path models with subject-specific academic self-efficacy and achievement goals as mediators. The model displayed satisfactory fit to the data in both subject domains,  $\chi^2(7, N = 304) = 13.927, p > .05$  (TLI = .931, CFI = .989, RMSEA = .057) in math, and  $\chi^2(7, N = 304) = 10.980, p > .05$  (TLI = .959, CFI = .994, RMSEA = .043) in English. Figure 3 presents statistically significant paths at  $p < .05$  from these models, with coefficients from the math model to the left of the slash and those from the English model to the right of the slash.

**Paths from perfectionism to motivation variables.** All of our hypotheses regarding the relationships of each perfectionism variable with academic self-efficacy and achievement goals received support. Specifically, self-oriented perfectionism related positively to academic self-efficacy ( $\beta_s = .46$  in math and  $.41$  in English), a mastery goal ( $\beta_s = .34$  in math and  $.35$  in English), and a performance-approach goal ( $\beta_s = .33$  in math and  $.21$  in English) in the respective subject. Socially prescribed perfectionism related positively to performance-approach ( $\beta_s = .34$  in math and  $.47$  in English) and performance-avoidance goals ( $\beta_s = .32$  in math and  $.44$  in English). The significant bivariate correlation between self-oriented perfectionism and a performance-avoidance goal and that between socially prescribed perfectionism and a mastery goal were no longer observed when only the unique variance in perfectionism was considered.

Two consistent mediation paths between self-oriented perfectionism and achievement goals by academic self-efficacy emerged. One involved a mastery goal ( $z = 2.87, p < .01$  in math and  $z = 3.18, p < .01$  in English) and the other a performance-avoidance goal ( $z = -3.36, p < .001$  in math and  $z = -3.27, p < .01$  in English). Self-oriented perfectionism related positively to academic self-efficacy in both subjects, while academic self-efficacy in turn related positively to a mastery goal ( $\beta_s = .24$  in math and  $.32$  in English) and negatively to a performance-avoidance goal in the domain ( $\beta_s = -.30$  in math and  $-.32$  in English). Table 3 presents estimates of indirect effects from the math model, and Table 4 presents those from the English model, along with results from the Sobel tests.

**Paths from self-oriented perfectionism to outcome variables.** Academic self-efficacy and a mastery goal mediated the significant negative path from self-oriented perfectionism to acceptability of cheating. Whereas pursuing a mastery goal alone sufficed as a mediator in this relationship, feeling self-efficacious did not. A sense of self-efficacy had to be coupled with a mastery goal in the subject domain, which then related negatively to acceptability of cheating ( $\beta_s = -.27$  in math and  $-.32$  in English). However, although the individual paths linking self-oriented perfectionism to self-efficacy, self-efficacy to a mastery goal, and a mastery goal to acceptability of cheating were all statistically significant in both subjects, the total indirect effects from self-oriented perfectionism to acceptability of cheating were not statistically significant, as determined by the bootstrapping method. Only the paths linking math self-efficacy to acceptability of cheating via a mastery goal in math proved significant ( $z = -2.17, p < .05$ ).

The direct negative path between self-oriented perfectionism and academic procrastination remained significant and negative even in the presence of intervening motivational variables. Still,

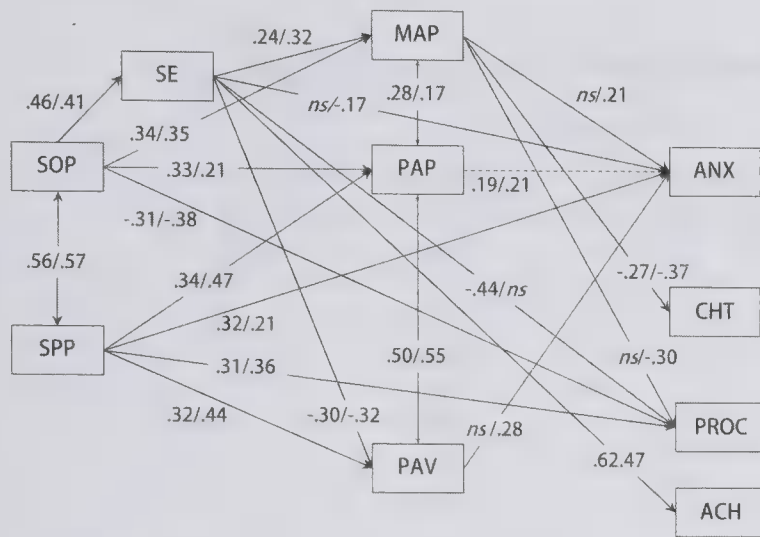


Figure 3. Final path model. Only statistically significant paths at  $p < .05$  are presented, except for the dotted path with  $p < .06$ . Error terms are omitted for clarity. Path coefficients in math are presented to the left of the slash; those in English to the right of the slash. SOP = self-oriented perfectionism; SPP = socially prescribed perfectionism; SE = academic self-efficacy; MAP = mastery goal; PAP = performance-approach goal; PAV = performance-avoidance goal; CHT = acceptability of cheating; ANX = test anxiety; PROC = academic procrastination; ACH = achievement scores.

the coefficients were in substantially reduced magnitude ( $\beta$ s =  $-.31$  in math and  $-.38$  in English) from the coefficient from the direct path model ( $\beta = -.66$ ), suggesting mediation effects. The path was partially mediated by academic self-efficacy in math ( $z = -3.73$ ,  $p < .001$ ) and by either a mastery goal alone ( $z = -2.10$ ,  $p < .05$ ) or by academic self-efficacy and a mastery goal together ( $z = -2.06$ ,  $p < .05$ ) in English. In math, self-oriented perfectionism related positively to academic self-efficacy, which related negatively to academic procrastination ( $\beta = -.44$ ). In English, self-oriented perfectionism related to a mastery goal either directly ( $\beta = .35$ ) or via academic self-efficacy ( $\beta = .32$ ). A mastery goal then related negatively to academic procrastination ( $\beta = -.30$ ).

The significant positive path from self-oriented perfectionism to achievement previously observed in the direct path model was fully mediated by subject-specific academic self-efficacy in both math ( $z = 4.74$ ,  $p < .001$ ) and English ( $z = 3.78$ ,  $p < .001$ ). Self-oriented perfectionism related positively to academic self-efficacy in the subject, which positively predicted achievement in both math ( $\beta = .62$ ) in English ( $\beta = .47$ ).

**Paths from socially prescribed perfectionism to outcome variables.** Whereas the paths from self-oriented perfectionism to outcome variables were largely mediated by academic self-efficacy and the two approach-oriented achievement goals, those from socially prescribed perfectionism were not. Two of the three direct paths from socially prescribed perfectionism to outcome variables previously observed in the direct path model remained significant and in comparable magnitude in the mediation models. Specifically, socially prescribed perfectionism was a direct positive predictor of test anxiety ( $\beta$ s =  $.32$  in math and  $.21$  in English) and academic procrastination ( $\beta$ s =  $.31$  in math and  $.36$  in English) in both subject domains. The only significant partial mediation was between socially prescribed perfectionism and test anxiety

in English ( $z = 2.26$ ,  $p < .05$ ). Socially prescribed perfectionism related positively to a performance-avoidance goal in English ( $\beta = .44$ ), which related positively to test anxiety ( $\beta = .28$ ).

## Discussion

Students come into class armed with not only motivation and prior knowledge but also family background, developmental and socialization history, and personality characteristics. It is important to learn how these diverse factors all come into play in achievement settings, at least what the salient patterns are among major variables, to understand the “whole” student. It is for this reason that we were interested in the role of perfectionism in academic contexts in the first place. The present results once again demonstrated that the effects of stable personality dispositions, such as perfectionism on academic outcomes, although not trivial, do get mediated by students’ motivational beliefs in specific subject domains.

Our primary purpose in this research was twofold. First, we tried to ascertain the dimensional nature of perfectionism, so as to help future research with this personality trait with potentially weighty consequences for students in achievement settings. Whereas researchers seldom question the maladaptive nature of socially prescribed perfectionism, they have not been able to reach a firm conclusion regarding the adaptive nature of self-oriented perfectionism (Stoeber et al., 2009). We reasoned that dimensionality of perfectionism would play out most vividly when assessed in reference to typical learning situations, due to the ongoing competition, imminent possibilities of failure, and ambiguous definitions of success inherent in them.

Second, we wanted to test once again the importance of academic motivation in assisting learners with adaptive as well as maladaptive dispositions to adjust and function better in specific achievement situations. Because personality traits predispose learners to certain motivational tendencies, academic motivation likely mediates the processes linking perfectionism to achievement-related outcomes (Mills & Blankstein, 2000). Further, because self-oriented and socially prescribed perfectionism differ with respect to who initiates and maintains control over goals and standards, learners high in each type of perfectionism inevitably generate different responses toward identical challenges and setbacks and are expected to conclude the same achievement episodes differently by following disparate motivational paths.

The results were largely consistent with our hypotheses. Self-oriented perfectionism related positively to academic achievement and negatively to acceptability of cheating and academic procrastination in achievement settings. It did not link significantly to test anxiety. Socially prescribed perfectionism, in contrast, related positively to test anxiety, acceptability of cheating, and academic procrastination but did not link significantly to academic achievement. Many of the significant paths from perfectionism to outcomes were mediated by domain-specific motivation. The paths from self-oriented perfectionism to outcomes were mediated by academic self-efficacy and a mastery goal in the domain, while those from socially prescribed perfectionism were mediated by a performance-avoidance goal. Nonetheless, the direct paths from the two perfectionism dimensions to academic procrastination and that from socially prescribed perfectionism to test anxiety re-



Table 3  
Standardized Total, Direct, and Indirect Effects in the Math Model

Independent variable	Mediator variable	Dependent variable	Standardized estimate of indirect effects			Sobel test <i>z</i>
			Total	Direct	Indirect	
SOP	→ SE	→ MAP	.45**	.34**	.11**	2.87**
		→ MAP				
		→ PAP	.31**	.33**	-.02	
	→ SE	→ PAV	.04	.18 <sup>†</sup>	-.14**	-3.36***
		→ PAV				
		→ ANX	.01	-.04	.05	
	→ SE	→ CHT	-.26*	-.16	-.10	-3.73***
		→ PROC	-.57**	-.31 <sup>†</sup>	-.26**	
		→ PROC				
	→ SE	→ ACH	.23*	-.06	.29**	4.74***
		→ ACH				
		→ ACH				
SPP	→ PAP	→ MAP	-.02	.01	-.04	1.68 <sup>†</sup>
		→ PAP	.35**	.34**	.01	
		→ PAV	.37**	.32*	.04	
		→ ANX	.44**	.32*	.12**	
	→ PAP	→ ANX				
		→ CHT	.17	.08	.09	
		→ PROC	.35**	.31*	.04	
		→ ACH	.02	.12	-.10	
SE	→ MAP	→ ANX	-.11	-.07	-.04	-2.17*
		→ CHT	-.15 <sup>†</sup>	-.07	-.08*	
		→ CHT				
		→ PROC	-.45**	-.44**	-.02	
		→ PROC				
		→ ACH	.63**	.62**	.01	

Note. SOP = self-oriented perfectionism; SPP = socially prescribed perfectionism; SE = academic self-efficacy; MAP = mastery goal; PAV = performance-avoidance goal; PAP = performance-approach goal; ANX = test anxiety; CHT = acceptability of cheating; PROC = academic procrastination; ACH = achievement scores. Based on 1,000 bootstrap samples.

<sup>†</sup>  $p < .10$ . \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

mained significant, even in the presence of the intervening motivation variables.

### Self-Oriented as Adaptive Perfectionism and Socially Prescribed as Maladaptive Perfectionism

Consistent with previous reports (Flett et al., 1994, 1995; Hewitt et al., 2002; Mills & Blankstein, 2000; Verner-Filion & Gaudreau, 2010), self-oriented perfectionism and socially prescribed perfectionism correlated with each other yet were clearly distinguishable dimensions for our Korean middle school participants. The correlation between the two perfectionism dimensions, however, was noticeably larger than what has been typically observed in the literature. The strong correlation between self-oriented and socially prescribed perfectionism seems to indicate that the Korean adolescents, with a strong desire to meet the extremely high standards that their teachers and parents set for them, also tended to set similarly high standards for themselves and strove to achieve those perfectionistic standards.

Interdependent self-construal could accentuate socially prescribed perfectionism as well as strengthen alliance between the two perfectionism dimensions. The desire to please significant others in the social network that one identifies with, is a strong source of motivation for individuals in collectivistic cultures with interdependent self-construal (Heine, 2001; Markus & Kitayama, 1991; Oishi & Diener, 2001). Because conformity is a virtue

(Markus & Kitayama, 1994), they would ascribe high value to what their in-group members consider important. Korean students, presumably with stronger interdependent self-construal compared to students in Western cultures, could more likely approve the goals and standards valued by parents and teachers and internalize them as their own. This conjecture should be formally tested in future research, however, as the mean score of socially prescribed perfectionism in this study was not particularly high.

In previous research, self-oriented perfectionism has frequently demonstrated both positive and negative characteristics, correlating positively with variables as varied as depression, test anxiety, self-efficacy, and intrinsic and extrinsic motivation (Hewitt et al., 2002; Mills & Blankstein, 2000; Stoeber et al., 2009). For the Korean adolescents participating in this study, in comparison, self-oriented perfectionism consistently emerged as a positive predictor of adaptive variables, including academic self-efficacy and achievement, and a negative predictor of maladaptive variables, including acceptability of cheating and academic procrastination. Socially prescribed perfectionism primarily functioned as a positive predictor of maladaptive variables. As we conjectured and consistent with Mills and Blankstein (2000), self-oriented perfectionism correlated positively with test anxiety, but when the covariance with socially prescribed perfectionism was controlled for, the relationship was no longer significant. Self-oriented perfectionism thus appears to be an adaptive characteristic, while socially

Table 4  
Standardized Total, Direct, and Indirect Effects in the English Model

Independent variable	Mediator variable	Dependent variable	Standardized estimate of indirect effects			Sobel test <i>z</i>
			Total	Direct	Indirect	
SOP	→ SE	→ MAP	.49**	.35**	.13**	3.18**
		→ PAP	.24*	.21 <sup>†</sup>	.03	
		→ PAV	.03	.16	-.13**	
	→ SE	→ PAV				-3.27**
		→ ANX	.00	-.09	.09	
		→ CHT	-.27*	-.13	-.14 <sup>†</sup>	
		→ PROC	-.58**	-.38*	-.20*	
	→ MAP	→ PROC				-2.10*
	→ SE → MAP	→ PROC				-2.06*
	→ SE	→ ACH	.25**	-.02	.27**	3.78***
SPP	→ SE	→ ACH				
		→ MAP	-.02	.02	-.04	
		→ PAP	.46**	.47**	-.01	
		→ PAV	.48**	.44**	.04	
		→ ANX	.45**	.21 <sup>†</sup>	.25**	
	→ PAV	→ ANX				2.26*
		→ CHT	.19	.07	.12	
		→ PROC	.36**	.36*	.00	
SE		→ ACH	-.04	-.02	-.02	
		→ ANX	-.17*	-.17*	-.01	
		→ CHT	-.13	-.02	-.11	
		→ PROC	-.21*	-.12	-.08	
		→ ACH	.52**	.47**	.06	

Note. SOP = self-oriented perfectionism; SPP = socially prescribed perfectionism; SE = academic self-efficacy; MAP = mastery goal; PAV = performance-avoidance goal; PAP = performance-approach goal; ANX = test anxiety; CHT = acceptability of cheating; PROC = academic procrastination; ACH = achievement scores. Based on 1,000 bootstrap samples.

<sup>†</sup>  $p < .10$ . \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

prescribed perfectionism a maladaptive characteristic, to possess in the academic domain—at least for the current sample of Korean middle school students.

The nature of direct paths from either type of perfectionism to subject-specific motivation was consistent across math and English. Self-oriented perfectionism related positively to academic self-efficacy, a mastery goal, and a performance-approach goal, while socially prescribed perfectionism did not relate to academic self-efficacy and instead related positively to performance-approach and performance-avoidance goals. Consistent with our hypothesis, the relationships of perfectionism with outcome variables were mediated by academic self-efficacy and achievement goals in the subject domain. The pattern of mediation also generally stayed the same, regardless of whether motivation in math or that in English was examined. This is strong evidence that present findings are not confined to particular subject matter domains.

Even so, there was a difference in the extent to which motivation mediated the effects of perfectionism. All paths from self-oriented perfectionism to outcome variables were mediated, either fully or partially, by motivation variables. Those from socially prescribed perfectionism, in contrast, were not as effectively mediated by the same variables. Two of the three direct paths to outcome variables remained significant and strong, even after the subject-specific motivation variables entered the equation. The present results thus suggest that self-oriented perfectionism works largely through motivation of learners in specific achievement contexts, whereas socially

prescribed perfectionism works more directly on achievement-related outcomes.

It is worth noting that the four variables to which socially prescribed perfectionism linked directly and consistently across the two subject domains—test anxiety, academic procrastination, a performance-approach goal, and a performance-avoidance goal—as well as the socially prescribed perfectionism itself, are correlates of fear of failure (Elliot & Church, 1997; Speirs Neumeister, 2004; Steel, 2007). This finding implies that fear of failure may be particularly resistant to contextual influences. Bong (2001) offered a similar conjecture when performance-approach and performance-avoidance goals of Korean middle and high school students displayed noticeably stronger correlations across multiple subject matter domains than did other motivation constructs. Compared to non-Asian students, Asian students also report stronger fear of failure, which explains their motivation in a specific achievement context better than do other constructs such as self-efficacy or effort attribution (Eaton & Dembo, 1997). Accordingly, it is possible that the relationships of socially prescribed perfectionism could be better mediated by motivation variables among non-Asian learners.

Of the direct and indirect paths from perfectionism to maladaptive outcomes included in this study, academic procrastination, in particular, clearly epitomizes the contrasting nature of the two perfectionism dimensions. As students expressed stronger self-oriented perfectionism, they were less likely to engage in academic



procrastination. As they expressed stronger socially prescribed perfectionism, on the contrary, they more frequently engaged in procrastinating behaviors. Socially prescribed perfectionism consistently demonstrates a strong positive correlation with fear of negative evaluation (Flett, Hewitt, & De Rosa, 1996; Hewitt & Flett, 1991). Also, compared to self-oriented perfectionism, which correlates positively with commitment to perfect "performance" at school, socially prescribed perfectionism correlates positively with commitment to perfect "relationships" with significant others (Flett et al., 1995). Students with socially prescribed perfectionism, finding it difficult to satisfy the perfectionistic standards of others, could resort to academic procrastination as a desperate means to delay unfavorable judgments by, and damaged relationships with, parents and teachers.

Milgram, Sroloff, and Rosenbaum (1988) provided yet another interesting account of the strong perfectionism-procrastination link. They argued that when students feel that parents, teachers, and other powerful adult figures impose certain tasks on them, they may show greater procrastination as an expression of covert negativism. Covert negativism is a type of motivation that represents an indirect display of hostility and passive retaliation toward authority figures. Steel (2007) also discussed a possibility that rebellious individuals, especially young adolescents, procrastinate on tasks with externally imposed deadlines because they view these tasks to be highly aversive. Socially prescribed perfectionism, by definition, refers to the excessive strivings toward and concerns about fulfilling the difficult standards imposed upon them by significant others (Hewitt & Flett, 1991). Covert negativism seems to explain well the stronger tendency among socially prescribed perfectionists to procrastinate in academic situations, especially adolescents in collectivistic cultures who find it difficult to ignore the wishes of their parents.

### **Academic Self-Efficacy as a Positive Amplifier of Self-Oriented Perfectionism**

We examined the role of perfectionism in concert with academic self-efficacy and achievement goals, arguably the two most prominent constructs in contemporary academic motivation research. Students' self-efficacy beliefs, or subjective convictions to perform successfully in the given subject domain (Bandura, 1997; Schunk, 1991), were particularly effective in augmenting the adaptive aspects of self-oriented perfectionism. The positive direct path from self-oriented perfectionism to academic achievement and the negative direct paths to acceptability of cheating and academic procrastination were all significantly mediated by academic self-efficacy in the domain.

More specifically, self-oriented perfectionism related to stronger academic self-efficacy in the subject domain, whether it was math or English. Stronger academic self-efficacy, in turn, related to a stronger mastery goal and a weaker performance-avoidance goal in the subject areas. It also related to less academic procrastination and acceptability of cheating among students, directly or indirectly through a stronger mastery goal. Most of all, academic self-efficacy was the strongest positive predictor of achievement in both math and English, a finding that is now clearly established in the academic motivation literature (Bong & Skaalvik, 2003; Mulleton et al., 1991; Pajares, 1996; Schunk, 1991; Zimmerman, 2000).

Whereas academic self-efficacy as a mediator reinforced the adaptive functions of self-oriented perfectionism, it did not intervene between socially prescribed perfectionism with other variables. Korean middle school students expressed stronger convictions for successfully performing in the given subject domains as they expressed a stronger desire to achieve highly difficult goals but only when those goals were set by themselves. The desire to satisfy difficult goals set forth by others—socially prescribed perfectionism—did not demonstrate a systematic relationship with academic self-efficacy. The goals and standards set by self-oriented perfectionists, though they may be excessively demanding, thus appear to instill a sense of agency and perceived control in the individuals, which results in stronger convictions in their own capabilities for successfully attaining desired outcomes in the given domains.

Latham and Locke (1991) wrote, "Given sufficient ability, goal theory predicts a drop at high goal difficulty levels . . . if there is a large decrease in goal commitment" (p. 215). As discussed above, socially prescribed perfectionism correlated not with commitment to perfect performance at school but with commitment to perfect relationships with significant others (Flett et al., 1995). This suggests that socially prescribed perfectionists lack commitment to the goals of achieving the perfectionistic performance imposed on them. This lack of goal commitment would disrupt the relationships between socially prescribed perfectionism, self-efficacy, and performance. Whereas socially prescribed perfectionism does not appear to have direct implications for self-efficacy of students in the academic domain, it does appear to orient students toward particular types of achievement goals, which we discuss next.

### **Types of Perfectionism as Antecedents of Achievement Goals**

As hypothesized, self-oriented perfectionism linked positively to a mastery goal and a performance-approach goal. Socially prescribed perfectionism linked positively to a performance-approach goal and a performance-avoidance goal. The same pattern emerged in both math and English. A commonality between self-oriented perfectionism and the two approach-oriented achievement goals is having an achievement motive as an antecedent. Self-oriented perfectionism and a mastery goal also prompt standards-based, as opposed to comparison-based, competence evaluation. Socially prescribed perfectionism and the two performance-oriented goals have fear of failure as a common correlate. Competence appraisals in socially prescribed perfectionism are carried out against standards imposed by others, while those in performance goals are executed against criteria determined by others' performance (Elliot & Church, 1997; Elliot & McGregor, 2001; Speirs Neumeister, 2004; Van Yperen, 2006). It is hence not surprising that self-oriented perfectionists readily adopt a mastery goal, while socially prescribed perfectionists readily adopt performance goals in achievement situations.

For socially prescribed perfectionists, "others" are more than simply a source of comparison standards. Fear of unfavorable evaluation from others is a well-established correlate of socially prescribed perfectionism (Flett et al., 1996; Hewitt & Flett, 1991). In achievement situations, such fear could translate into concerns about proving competence to and concealing incompetence from



others. In this study, we assessed achievement goals with the items used in a study by Elliot and McGregor (2001). The performance-approach goal items focused exclusively on normative competence without any reference to concerns for ability validation, while the performance-avoidance goal items retained components of ability validation (e.g., "My fear of performing poorly in this class is often what motivates me"). Socially prescribed perfectionism still correlated strongly with both performance goals in both math and English. Had we used a different set of performance goal items that combines the normative competence and ability validation components (e.g., PALS; Midgley et al., 2000), even stronger relationships could have emerged between socially prescribed perfectionism and the two performance goals.

A recent debate in achievement goal research entails whether the concern for demonstrating and validating ability in front of others should be viewed as a legitimate and indispensable constituent of performance goals (Elliot & Murayama, 2008; Grant & Dweck, 2003). The present results do not speak directly to this question. Nevertheless, it deserves to note that students' desire to appear perfectly competent by satisfying the standards imposed on them by others significantly and consistently related to both performance-approach and performance-avoidance goals across two specific subject matter domains in this study. Further, the paths from socially prescribed perfectionism to performance-approach goals ( $\beta_s = .33$  and  $.21$ ) were comparable in strengths with those to performance-avoidance goals ( $\beta_s = .34$  and  $.47$ ).

If socially prescribed perfectionism is maladaptive, these results suggest that not only a performance-avoidance goal but also a performance-approach goal share its maladaptive characteristics. The positive correlations of performance-approach goals with performance-avoidance goals ( $r_s = .63$  and  $.59$ ) and test anxiety in this study ( $r_s = .50$  and  $.42$ ) support this conjecture. In fact, the potentially detrimental nature of a performance-approach goal, amidst its positive associations with performance indexes, has been repeatedly observed in previous studies with younger learners. Korean elementary and middle school students, for example, do not distinguish between performance-approach and performance-avoidance goals (Bong, Woo, & Shin, 2013) and, even when they do, their performance-approach goals correlate significantly with test anxiety (Bong, 2009) and predict help-seeking avoidance (Bong, 2008). Continued research on the makeup and function of the performance goal, therefore, seems warranted.

## Limitations

Several limitations of the present investigation should be noted. First, we assumed certain causal predominance among the variables included in our model according to theory and previous research. However, we measured the variables concurrently, except for the achievement scores that were collected after the surveys. A more accurate test of the mediating processes requires that presumed antecedents and consequents be assessed with a sufficient temporal interval.

Second, while we assessed motivation and achievement variables in reference to specific subject matter areas, outcome variables such as test anxiety, acceptability of cheating, and academic procrastination were assessed in reference to general learning situations. Because one of our primary interests was direct relationships between the two perfectionism dimensions and key out-

come variables, and because a difference in assessment specificity between constructs could hamper proper examination of their associations (Pajares & Miller, 1995), we decided to assess the outcome variables at a level most similar to that of perfectionism. Evidence that anxiety (Green, Martin, & Marsh, 2007), cheating (Burton, 1963), and procrastination (Milgram, Mey-Tal, & Levi-son, 1998) display strong cross-situational consistency aided our decision (but see Goetz, Frenzel, Pekrun, & Hall, 2006). Nevertheless, assessing all variables in the context of specific subject domains could disclose an interesting idiosyncrasy associated with subject matter learning, which we might have overlooked in this investigation.

Third, we used the Multidimensional Perfectionism Scale (MPS) by Hewitt and Flett (1991). This scale is by far the most frequently used one in the literature, and it has also been successfully translated and validated for Korean students in previous research (Seo & Synn, 2006). However, had we used a newer version specifically developed for the younger population, the Child-Adolescent Perfectionism Scale (CAPS; Flett, Hewitt, Boucher, Davidson, & Munro, 1997), the results could have been more accurate.

Fourth, the acceptability of cheating and academic procrastination scales demonstrated internal consistency estimates that were less than satisfactory. We were not too seriously concerned about the low reliability of these scales because (a) the acceptability of cheating scale had been associated with similar estimates of internal consistency in previous studies (Anderman et al., 1998; Murdock et al., 2004), (b) we only used reliable portions of the variance in the analysis, and (c) the relationships of the variables assessed with these scales with other variables were consistent with theory and previous findings. Nonetheless, the low reliability of these scales could have compromised integrity of the present findings to a certain degree.

## Contributions and Future Directions

Although perfectionism is a personality trait with strong motivational implications (Hewitt & Flett, 1991), only a limited number of studies to date have directly investigated how perfectionism relates to motivation and achievement in the academic domain (see Fletcher & Speirs Neumeister, 2012). Further, a majority of these studies stop at reporting correlations between perfectionism and other variables, without probing the potentially intricate mediation or moderation in their associations (Verner-Filion & Gaudreau, 2010). The present research fills this gap in the literature and documents relevance of multidimensional perfectionism for adolescents in achievement settings. Korean adolescents as young as middle school students differentiated self-oriented and socially prescribed dimensions of perfectionism. They also manifested a distinct pattern of motivation and achievement-related behavior, depending on the particular type of perfectionism they possessed. When distinguished from socially prescribed perfectionism, self-oriented perfectionism was more facilitative than disruptive for motivation and learning processes. The current investigation contributes to the literature by supporting the dimensional analysis of perfectionism and adding to the growing body of literature that suggests the relatively adaptive nature of self-oriented perfectionism (Stoeber et al., 2009).



More important, this study has demonstrated that motivation in specific academic contexts mediates the paths linking stable personality dispositions such as perfectionism to concrete affective, behavioral, and performance outcomes in the academic domain. In previous research on the role of perfectionism in students' academic functioning, investigators have typically examined the direct associations between perfectionism and outcomes without considering the intervening motivational processes (Bieling et al., 2003). When motivation variables were included, they were often assessed as motivation for school learning in general, and not as domain-specific motivational beliefs (Stoeber & Rambow, 2007; Verner-Filion & Gaudreau, 2010). However, such direct-path-only or general models would be an oversimplification of the complex interrelations among perfectionism, motivation, and outcomes. We tried to delineate part of this complexity by assessing some of the representative motivational constructs in reference to specific subject domains. By doing so, we were able to demonstrate that the manner with which each perfectionism dimension links to various outcomes depends, to a considerable degree, on students' self-efficacy beliefs and achievement goals in particular subject matter areas. We believe this is an important finding because, even if we cannot easily change the perfectionism in students, we can still improve the quality of the learning process they engage in by altering their domain-specific motivational beliefs.

Can we say, based on the present findings, that self-oriented perfectionism is truly an adaptive personality trait for students' academic functioning? The answer to this question depends on several conditions. Most of all, although self-oriented perfectionism played a positive role in this study, it is important to remember that it is not a pure form of achievement motivation that mutually excludes fear of failure or fear of negative evaluation from others (Flett et al., 1996; Frost et al., 1990; Hewitt & Flett, 1991; Speirs Neumeister, 2004). On the one hand, it represents the relentless propensity to demand a lot from oneself by setting high goals, which typically promotes intrinsic motivation, self-efficacy, effort, and persistence for attaining those goals. Under normal achievement situations where perceived stress or evaluative threat is not extreme, self-oriented perfectionism will activate approach-oriented motivation. On the other hand, it could turn maladaptive and function more similarly to socially prescribed perfectionism under extremely stressful situations.

When Hewitt et al. (2002) divided children into three groups by levels of achievement stress, for example, it was only the children with high and average levels of achievement or social stress for whom stronger self-oriented perfectionism resulted in greater depression and anxiety. For those children with low levels of achievement or social stress, self-oriented perfectionism did not show a significant relationship with any of these variables. The relationships of socially prescribed perfectionism with maladjustment symptoms did not depend on levels of stress. When individuals are under high stress, perceive strong evaluative threat, or need to perform high-stakes tasks that are of great importance, self-oriented perfectionism correlates with negative affect, depression, and anxiety (Frost & Marten, 1990; Hewitt, Mittelstaedt, & Wollert, 1989; Stoeber & Rambow, 2007), just like socially prescribed perfectionism does. We thus suggest that researchers and practitioners exercise due caution when interpreting findings related to multidimensional perfectionism, taking into account the known individual and situational moderators of perfectionism.

Finally, we suggest that socialization history and resultant dispositional characteristics may present another common ground on which different dimensions of perfectionism could link to specific motivational beliefs such as academic self-efficacy and achievement goals. Elliot and McGregor (2001) reported that person-focused negative feedback and conditional approval of mothers were antecedents of a performance-avoidance goal. Mothers' conditional approval was an antecedent of a performance-approach goal as well. Hollender (1965) described a similar socialization mechanism spawning perfectionism by stating, "Perfectionism most commonly develops in an insecure child who needs approval, acceptance and affection from parents who are difficult to please" (p. 103). Speirs Neumeister and Finch (2006) demonstrated that insecure attachment to parents was indeed an antecedent of both perfectionism dimensions, while others showed that socially prescribed perfectionism usually demonstrates considerably stronger correlations with parent-related variables such as parental expectations and parental criticism (Flett et al., 1995). Future research should explore the causal chain among socialization history, motive dispositions, and perfectionism of the children and their motivation in school.

## References

- Anderman, E. M., Griesinger, T., & Westerfield, G. (1998). Motivation and cheating during early adolescence. *Journal of Educational Psychology, 90*, 84–93. doi:10.1037/0022-0663.90.1.84
- Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*, 155–173. doi:10.1007/BF02294170
- Arbuckle, J. L. (2006). *AMOS 7.0 user's guide*. Chicago, IL: SPSS.
- Bandalos, D. L., Finney, S. J., & Geske, J. A. (2003). A model of statistics performance based on achievement goal theory. *Journal of Educational Psychology, 95*, 604–616. doi:10.1037/0022-0663.95.3.604
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*, 191–215. doi:10.1037/0033-295X.84.2.191
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Freeman.
- Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology, 41*, 586–598. doi:10.1037/0022-3514.41.3.586
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246. doi:10.1037/0033-2909.107.2.238
- Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74*, 137–143. doi:10.1007/s11336-008-9100-1
- Bieling, P. J., Israeli, A., Smith, J., & Antony, M. M. (2003). Making the grade: The behavioral consequences of perfectionism in the classroom. *Personality and Individual Differences, 35*, 163–178. doi:10.1016/S0191-8869(02)00173-3
- Bong, M. (1997). Generality of academic self-efficacy judgments: Evidence of hierarchical relations. *Journal of Educational Psychology, 89*, 696–709. doi:10.1037/0022-0663.89.4.696
- Bong, M. (2001). Between- and within-domain relations of academic motivation among middle and high school students: Self-efficacy, task-value, and achievement goals. *Journal of Educational Psychology, 93*, 23–34. doi:10.1037/0022-0663.93.1.23
- Bong, M. (2005). Within-grade changes in Korean girls' motivation and perceptions of the learning environment across domains and achieve-

- ment levels. *Journal of Educational Psychology*, 97, 656–672. doi:10.1037/0022-0663.97.4.656
- Bong, M. (2008). Effects of parent–child relationships and classroom goal structures on motivation, help-seeking avoidance, and cheating. *Journal of Experimental Education*, 76, 191–217. doi:10.3200/JEXE.76.2.191-217
- Bong, M. (2009). Age-related differences in achievement goal differentiation. *Journal of Educational Psychology*, 101, 879–896. doi:10.1037/a0015945
- Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, 15, 1–40. doi:10.1023/A:1021302408382
- Bong, M., Woo, Y., & Shin, J. (2013). Do students distinguish between different types of performance goals? *Journal of Experimental Education*, 81, 464–489. doi:10.1080/00220973.2012.745464
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & S. J. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.
- Burns, L. R., Dittmann, K., Nguyen, N.-L., & Mitchelson, J. K. (2000). Academic procrastination, perfectionism, and control: Associations with vigilant and avoidant coping. *Journal of Social Behavior & Personality*, 15, 35–46.
- Burton, R. V. (1963). Generality of honesty reconsidered. *Psychological Review*, 70, 481–499. doi:10.1037/h0047594
- Castro, J. R., & Rice, K. G. (2003). Perfectionism and ethnicity: Implications for depressive symptoms and self-reported academic achievement. *Cultural Diversity and Ethnic Minority Psychology*, 9, 64–78. doi:10.1037/1099-9809.9.1.64
- Cury, F., Elliot, A. J., Da Fonseca, D., & Moller, A. C. (2006). The social-cognitive model of achievement motivation and the 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology*, 90, 666–679. doi:10.1037/0022-3514.90.4.666
- Daniels, L. M., Stupnisky, R. H., Pekrun, R., Haynes, T. L., Perry, R. P., & Newall, N. E. (2009). A longitudinal analysis of achievement goals: From affective antecedents to emotional effects and achievement outcomes. *Journal of Educational Psychology*, 101, 948–963. doi:10.1037/a0016096
- Duncan, T. G., & McKeachie, W. J. (2005). The making of the Motivated Strategies for Learning Questionnaire. *Educational Psychologist*, 40, 117–128. doi:10.1207/s15326985sep4002\_6
- Dunkley, D. M., Zuroff, D. C., & Blankstein, K. R. (2003). Self-critical perfectionism and daily affect: Dispositional and situational influences on stress and coping. *Journal of Personality and Social Psychology*, 84, 234–252. doi:10.1037/0022-3514.84.1.234
- Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, 95, 256–273. doi:10.1037/0033-295X.95.2.256
- Eaton, M. J., & Dembo, M. H. (1997). Differences in the motivational beliefs of Asian American and non-Asian students. *Journal of Educational Psychology*, 89, 433–440. doi:10.1037/0022-0663.89.3.433
- Einstein, D. A., Lovibond, P. F., & Gaston, J. E. (2000). Relationship between perfectionism and emotional symptoms in an adolescent sample. *Australian Journal of Psychology*, 52, 89–93. doi:10.1080/00049530008255373
- Elliot, A. J., & Church, M. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology*, 72, 218–232. doi:10.1037/0022-3514.72.1.218
- Elliot, A. J., & McGregor, H. A. (2001). A 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology*, 80, 501–519. doi:10.1037/0022-3514.80.3.501
- Elliot, A. J., & Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology*, 100, 613–628. doi:10.1037/0022-0663.100.3.613
- Fletcher, K. L., Shim, S. S., & Wang, C. (2012). Perfectionistic concerns mediate the relationship between psychologically controlling parenting and achievement goal orientations. *Personality and Individual Differences*, 52, 876–881. doi:10.1016/j.paid.2012.02.001
- Fletcher, K. L., & Speirs Neumeister, K. L. (2012). Research on perfectionism and achievement motivation: Implications for gifted students. *Psychology in the Schools*, 49, 668–677. doi:10.1002/pits.21623
- Flett, G. L., Blankstein, K. R., Hewitt, P. L., & Koledin, S. (1992). Components of perfectionism and procrastination in college students. *Social Behavior and Personality*, 20, 85–94. doi:10.2224/sbp.1992.20.2.85
- Flett, G. L., Hewitt, P. L., Boucher, D. J., Davidson, L. A., & Munro, Y. (1997). *The child-adolescent perfectionism scale: Development, validation, and association with adjustment*. Unpublished manuscript, Department of Psychology, York University, North York, Ontario, Canada.
- Flett, G. L., Hewitt, P. L., & De Rosa, T. (1996). Dimensions of perfectionism, psychosocial adjustment, and social skills. *Personality and Individual Differences*, 20, 143–150. doi:10.1016/0191-8869(95)00170-0
- Flett, G. L., Hewitt, P. L., & Dyck, D. G. (1989). Self-oriented perfectionism, neuroticism, and anxiety. *Personality and Individual Differences*, 10, 731–735. doi:10.1016/0191-8869(89)90119-0
- Flett, G. L., Hewitt, P. L., Endler, N. S., & Tassone, C. (1994). Perfectionism and components of state and trait anxiety. *Current Psychology*, 13, 326–350. doi:10.1007/BF02686891
- Flett, G. L., Sawatzky, D. L., & Hewitt, P. L. (1995). Dimensions of perfectionism and goal commitment: A further comparison of two perfectionism measures. *Journal of Psychopathology and Behavioral Assessment*, 17, 111–124. doi:10.1007/BF02229013
- Fritzsche, B. A., Young, B. R., & Hickson, K. C. (2003). Individual differences in academic procrastination tendency and writing success. *Personality and Individual Differences*, 35, 1549–1557. doi:10.1016/S0191-8869(02)00369-0
- Frost, R. O., & Marten, P. A. (1990). Perfectionism and evaluative threat. *Cognitive Therapy and Research*, 14, 559–572. doi:10.1007/BF01173364
- Frost, R. O., Marten, P., Lahart, C., & Rosenblate, R. (1990). The dimensions of perfectionism. *Cognitive Therapy and Research*, 14, 449–468. doi:10.1007/BF01172967
- Goetz, T., Frenzel, A. C., Pekrun, R., & Hall, N. C. (2006). The domain specificity of academic emotional experiences. *Journal of Experimental Education*, 75, 5–29. doi:10.3200/JEXE.75.1.5-29
- Grant, H., & Dweck, C. S. (2003). Clarifying achievement goals and their impact. *Journal of Personality and Social Psychology*, 85, 541–553. doi:10.1037/0022-3514.85.3.541
- Green, J., Martin, A. J., & Marsh, H. W. (2007). Motivation and engagement in English, mathematics and science high school subjects: Towards an understanding of multidimensional domain specificity. *Learning and Individual Differences*, 17, 269–279. doi:10.1016/j.lindif.2006.12.003
- Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and Advances*. Baltimore, MD: Johns Hopkins University Press.
- Heine, S. J. (2001). Self as cultural product: An examination of East Asian and North American selves. *Journal of Personality*, 69, 881–906. doi:10.1111/1467-6494.696168
- Hewitt, P. L., Caelian, C. F., Flett, G. L., Sherry, S. B., Collins, L., & Flynn, C. A. (2002). Perfectionism in children: Associations with depression, anxiety, and anger. *Personality and Individual Differences*, 32, 1049–1061. doi:10.1016/S0191-8869(01)00109-X
- Hewitt, P. L., & Flett, G. L. (1991). Perfectionism in the self and social contexts: Conceptualization, assessment, and association with psychopathology. *Journal of Personality and Social Psychology*, 60, 456–470. doi:10.1037/0022-3514.60.3.456



- Hewitt, P. L., Mittelstaedt, W., & Wollert, R. (1989). Validation of a measure of perfectionism. *Journal of Personality Assessment*, 53, 133–144. doi:10.1207/s15327752jpa5301\_14
- Hollender, M. H. (1965). Perfectionism. *Comprehensive Psychiatry*, 6, 94–103. doi:10.1016/S0010-440X(65)80016-5
- Howell, A. J., Watson, D. C., Powell, R. A., & Buro, K. (2006). Academic procrastination: The pattern and correlates of behavioural postponement. *Personality and Individual Differences*, 40, 1519–1530. doi:10.1016/j.paid.2005.11.023
- Hu, L., & Bentler, P. M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55. doi:10.1080/10705519909540118
- Hulleman, C. S., Schrager, S. M., Bodmann, S. M., & Harackiewicz, J. M. (2010). A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? *Psychological Bulletin*, 136, 422–449. doi:10.1037/a0018947
- Hwang, A. (2010). *Relationships of social goals with personality, perceived school goal structures, achievement goals, and academic outcomes* (Unpublished master's thesis). Korea University, Seoul, Korea.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.
- Latham, G. P., & Locke, E. A. (1991). Self-regulation through goal setting. *Organizational Behavior and Human Decision Processes*, 50, 212–247. doi:10.1016/0749-5978(91)90021-K
- Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57, 705–717. doi:10.1037/0003-066X.57.9.705
- Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98, 224–253. doi:10.1037/0033-295X.98.2.224
- Markus, H. R., & Kitayama, S. (1994). A collective fear of the collective: Implications for selves and theories of selves. *Personality and Social Psychology Bulletin*, 20, 568–579. doi:10.1177/0146167294205013
- Marsh, H. W., Byrne, B. M., & Shavelson, R. J. (1988). A multifaceted academic self-concept: Its hierarchical structure and its relation to academic achievement. *Journal of Educational Psychology*, 80, 366–380. doi:10.1037/0022-0663.80.3.366
- Middleton, M. J., & Midgley, C. (1997). Avoiding the demonstration of lack of ability: An underexplored aspect of goal theory. *Journal of Educational Psychology*, 89, 710–718. doi:10.1037/0022-0663.89.4.710
- Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., . . . Urdan, T. (2000). *Manual for the Patterns of Adaptive Learning Scales (PALS)*. Ann Arbor, MI: University of Michigan.
- Milgram, N., Marshevsky, S., & Sadeh, C. (1995). Correlates of academic procrastination: Discomfort task aversiveness and task capability. *Journal of Psychology: Interdisciplinary and Applied*, 129, 145–155. doi:10.1080/00223980.1995.9914954
- Milgram, N., Mey-Tal, G., & Levison, Y. (1998). Procrastination, generalized or specific, in college students and their parents. *Personality and Individual Differences*, 25, 297–316. doi:10.1016/S0191-8869(98)00044-0
- Milgram, N. A., Sroloff, B., & Rosenbaum, M. (1988). The procrastination of everyday life. *Journal of Research in Personality*, 22, 197–212. doi:10.1016/0092-6566(88)90015-3
- Mills, J. S., & Blankstein, K. R. (2000). Perfectionism, intrinsic vs. extrinsic motivation, and motivated strategies for learning: A multidimensional analysis of university students. *Personality and Individual Differences*, 29, 1191–1204. doi:10.1016/S0191-8869(00)00003-9
- Miquelon, P., Vallerand, R. J., Grouzet, F. M. E., & Cardinal, G. (2005). Perfectionism, academic motivation, and psychological adjustment: An integrative model. *Personality and Social Psychology Bulletin*, 31, 913–924. doi:10.1177/0146167204272298
- Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology*, 38, 30–38. doi:10.1037/0022-0167.38.1.30
- Murdock, T. B., Miller, A., & Kohlhardt, J. (2004). Effects of classroom context variables on high school students' judgments of the acceptability and likelihood of cheating. *Journal of Educational Psychology*, 96, 765–777. doi:10.1037/0022-0663.96.4.765
- Nathanson, C., Paulhus, D. L., & Williams, K. M. (2006). Predictors of a behavioral measure of scholastic cheating: Personality and competence but not demographics. *Contemporary Educational Psychology*, 31, 97–122. doi:10.1016/j.cedpsych.2005.03.001
- Oishi, S., & Diener, E. (2001). Goals, culture, and subjective well-being. *Personality and Social Psychology Bulletin*, 27, 1674–1682. doi:10.1177/01461672012712010
- Okagaki, L., & Frensch, P. A. (1998). Parenting and children's school achievement: A multiethnic perspective. *American Educational Research Journal*, 35, 123–144. doi:10.3102/00028312035001123
- Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, 66, 543–578. doi:10.3102/00346543066004543
- Pajares, F., & Miller, M. D. (1995). Mathematics self-efficacy and mathematics performances: The need for specificity of assessment. *Journal of Counseling Psychology*, 42, 190–198. doi:10.1037/0022-0167.42.2.190
- Park, Y. S., & Kim, U. (2006). Family, parent-child relationship, and academic achievement in Korea: Indigenous, cultural, and psychological analysis. In U. Kim, K. Yang, & K. Hwang (Eds.), *Indigenous and cultural psychology: Understanding people in context* (pp. 421–443). New York, NY: Springer Science + Business Media.
- Pekrun, R., Elliot, A. J., & Maier, M. A. (2006). Achievement goals and discrete achievement emotions: A theoretical model and prospective test. *Journal of Educational Psychology*, 98, 583–597. doi:10.1037/0022-0663.98.3.583
- Pintrich, P. R. (2000). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology*, 92, 544–555. doi:10.1037/0022-0663.92.3.544
- Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82, 33–40. doi:10.1037/0022-0663.82.1.33
- Pintrich, P. R., Smith, D., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement*, 53, 801–813. doi:10.1177/0013164493053003024
- Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, 21, 173–184. doi:10.1177/01466216970212006
- Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy*, 35, 299–331. doi:10.1016/S0005-7894(04)80041-8
- Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist*, 26, 207–231.
- Schunk, D. H. (1996). Goal and self-evaluative influences during children's cognitive skill learning. *American Educational Research Journal*, 33, 359–382. doi:10.3102/00028312033002359
- Schunk, D. H., & Ertmer, P. A. (1999). Self-regulatory processes during computer skill acquisition: Goal and self-evaluative influences. *Journal of Educational Psychology*, 91, 251–260. doi:10.1037/0022-0663.91.2.251
- Schunk, D. H., & Pajares, F. (2005). Competence perceptions and academic functioning. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 85–104). New York, NY: Guilford Press.
- Senko, C., & Harackiewicz, J. M. (2005). Achievement goals, task performance, and interest: Why perceived goal difficulty matters. *Personality*

- and *Social Psychology Bulletin*, 31, 1739–1753. doi:10.1177/0146167205281128
- Seo, E. H. (2008). Self-efficacy as a mediator in the relationship between self-oriented perfectionism and academic procrastination. *Social Behavior and Personality*, 36, 753–764. doi:10.2224/sbp.2008.36.6.753
- Seo, E. H., & Synn, M. H. (2006). The development and effect of a treatment program on academic procrastination according to perfectionism styles. *Korean Journal of Educational Research*, 44, 161–188.
- Sideridis, G. D. (2005). Goal orientation, academic achievement, and depression: Evidence in favor of a revised goal theory framework. *Journal of Educational Psychology*, 97, 366–375. doi:10.1037/0022-0663.97.3.366
- Solomon, L. J., & Rothblum, E. D. (1984). Academic procrastination: Frequency and cognitive-behavioral correlates. *Journal of Counseling Psychology*, 31, 503–509. doi:10.1037/0022-0167.31.4.503
- Speirs Neumeister, K. L. (2004). Understanding the relationship between perfectionism and achievement motivation in gifted college students. *Gifted Child Quarterly*, 48, 219–231. doi:10.1177/001698620404800306
- Speirs Neumeister, K. L., & Finch, H. (2006). Perfectionism in high-ability students: Relational precursors and influences on achievement motivation. *Gifted Child Quarterly*, 50, 238–251. doi:10.1177/001698620605000304
- Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin*, 133, 65–94. doi:10.1037/0033-2909.133.1.65
- Stoeber, J., Feast, A. R., & Hayward, J. A. (2009). Self-oriented and socially prescribed perfectionism: Differential relationships with intrinsic and extrinsic motivation and test anxiety. *Personality and Individual Differences*, 47, 423–428. doi:10.1016/j.paid.2009.04.014
- Stoeber, J., & Rambow, A. (2007). Perfectionism in adolescent school students: Relations with motivation, achievement, and well-being. *Personality and Individual Differences*, 42, 1379–1389. doi:10.1016/j.paid.2006.10.015
- Synn, M. H., Park, S. H., & Seo, E. H. (2005). Women college students' time management and procrastination on college grades. *Korean Journal of Educational Research*, 43, 211–230.
- Trumpeter, N., Watson, P. J., & O'Leary, B. J. (2006). Factors within multidimensional perfectionism scales: Complexity of relationships with self-esteem, narcissism, self-control, and self-criticism. *Personality and Individual Differences*, 41, 849–860. doi:10.1016/j.paid.2006.03.014
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1–10. doi:10.1007/BF02291170
- Vansteenkiste, M., Smeets, S., Soenens, B., Lens, W., Matos, L., & Deci, E. L. (2010). Autonomous and controlled regulation of performance-approach goals: Their relations to perfectionism and educational outcomes. *Motivation and Emotion*, 34, 333–353. doi:10.1007/s11031-010-9188-3
- Van Yperen, N. W. (2006). A novel approach to assessing achievement goals in the context of the 2 × 2 framework: Identifying distinct profiles of individuals with different dominant achievement goals. *Personality and Social Psychology Bulletin*, 32, 1432–1445. doi:10.1177/0146167206292093
- Verner-Filion, J., & Gaudreau, P. (2010). From perfectionism to academic adjustment: The mediating role of achievement goals. *Personality and Individual Differences*, 49, 181–186. doi:10.1016/j.paid.2010.03.029
- Wolters, C. A. (2003). Understanding procrastination from a self-regulated learning perspective. *Journal of Educational Psychology*, 95, 179–187. doi:10.1037/0022-0663.95.1.179
- Wolters, C. A. (2004). Advancing achievement goal theory: Using goal structures and goal orientations to predict students' motivation, cognition, and achievement. *Journal of Educational Psychology*, 96, 236–250. doi:10.1037/0022-0663.96.2.236
- Zeidner, M. (1994). Personal and contextual determinants of coping and anxiety in an evaluative situation: A prospective study. *Personality and Individual Differences*, 16, 899–918. doi:10.1016/0191-8869(94)90234-8
- Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25, 82–91. doi:10.1006/ceps.1999.1016

Received June 8, 2012

Revision received December 16, 2013

Accepted December 20, 2013 ■



# The Contribution of Adolescent Effortful Control to Early Adult Educational Attainment

Marie-Hélène Véronneau  
Université du Québec à Montréal

Kristina Hiatt Racer  
University of Oregon

Gregory M. Fosco  
Pennsylvania State University

Thomas J. Dishion  
Arizona State University

Effortful control has been proposed as a set of neurocognitive competencies that is relevant to self-regulation and educational attainment (Posner & Rothbart, 2007). This study tested the hypothesis that a multiagent report of adolescents' effortful control (age 17) would be predictive of academic persistence and educational attainment (age 23–25), after controlling for other established predictors (family factors, problem behavior, grade-point average, and substance use). Participants were 997 students recruited in 6th grade from 3 urban public middle schools (53% males; 42.4% European American; 29.2% African American). Consistent with the hypothesis, the unique association of effortful control with future educational attainment was comparable in strength to that of parental education and students' past grade-point average, suggesting that effortful control contributes to this outcome above and beyond well-established predictors. Path coefficients were equivalent across gender and ethnicity (European Americans and African Americans). Effortful control appears to be a core feature of the self-regulatory competencies associated with achievement of educational success in early adulthood. These findings suggest that the promotion of self-regulation in general and effortful control in particular may be an important focus not only for resilience to stress and avoidance of problem behavior but also for growth in academic competence.

**Keywords:** educational attainment level, self-regulation, academic achievement, adolescence, family background

Education success and attainment is the clearest index of competence and success in modern Western societies. At the individual level, higher educational attainment predicts quality of life throughout adulthood, including employment status, income, psychological and physical health, well-being, and community involvement (Adams, 2002; Day & Newburger, 2002; Herzog, Franks, Markus, & Holm-

berg, 1998; Karvonen et al., 2007; McCaul, Donaldson, Coladarci, & Davis, 1992; Ross & Mirowsky, 2006; Tobiasz-Adamczyk, Bartoszewska, Brzyski, & Kopacz, 2007; Zhang, Huang, Ye, & Zeng, 2008). From a societal perspective, it is necessary to promote higher rates of secondary school completion, postsecondary technical training, and college and graduate training to meet current socioeconomic and demographic challenges. These challenges include an aging workforce, which requires training of replacement workers, the fast pace of technological progress, and market globalization (Organisation for Economic Co-Operation and Development, 2005). During recent decades, researchers have identified many correlates of students' educational attainment, but high rates of school dropout and low attendance of postsecondary education programs still represent significant costs to industrialized countries, including the United States (Belfield, Levin, & Brookings, 2007) and Canada (Kirby, 2009). Thus, key targets must be identified for future intervention efforts aiming to help students persevere through their formal schooling. The main objective of this study was to examine the role of effortful control, an understudied yet promising predictor of school persistence, and to determine whether this predictor remains important after other known predictors of educational attainment are accounted for.

## Predictors of Educational Attainment

Many aspects of students' family background and individual characteristics have been studied in the search for significant

---

This article was published Online First February 17, 2014.

Marie-Hélène Véronneau, Department of Psychology, Université du Québec à Montréal, Montréal, Quebec, Canada; Kristina Hiatt Racer, Child and Family Center, University of Oregon; Gregory M. Fosco, Department of Human Development and Family Studies, Pennsylvania State University; Thomas J. Dishion, Department of Psychology, Arizona State University.

Funding was provided by National Institute on Drug Abuse Grants DA07031 and DA13773 awarded to Thomas J. Dishion, by National Institute of Mental Health Grant K01MH082127 awarded to Kristina Hiatt Racer, and by a start-up grant for new professors-researchers from the Faculty of Human Sciences, Université du Québec à Montréal awarded to Marie-Hélène Véronneau. We are deeply grateful for the hard work of the Project Alliance staff, study families, and participating schools; without them, this study would not have been possible. Thanks to Cheryl Mikkola for editorial assistance in the preparation of this manuscript.

Correspondence concerning this article should be addressed to Marie-Hélène Véronneau, Université du Québec à Montréal, Département de Psychologie, C.P. 8888, Succursale Centre-Ville, Montréal (QC) H3C 3P8, Canada. E-mail: veronneau.marie-helene@uqam.ca

predictors of educational attainment. Family socioeconomic status (SES) and family processes are two major predictive family characteristics that have been examined in relationship to children's educational progression.

Family SES is a multifaceted concept that affects children's long-term educational outcomes in at least two ways. First, parental education plays an important role in children's educational progression. Parents with higher levels of education are more likely to encourage their children to pursue higher education and to have the resources to support this endeavor. As such, parents' level of educational attainment is a strong and consistent predictor of students' academic persistence as measured in early and middle adulthood (Dubow, Boxer, & Huesmann, 2009; Hardy et al., 1997; King, Meehan, Trim, & Chassin, 2006; Kristensen, Gravseth, & Bjerkedal, 2009; Marjoribanks, 2005; Taylor, Hurd, Seltzer, Greenberg, & Floyd, 2010), even after controlling for other significant indicators of family SES, including the value or ownership of their housing, family income, and the prestige of parents' occupation (Albrecht & Albrecht, 2011; Dubow et al., 2009; Kristensen et al., 2009; Melby, Conger, Fang, Wickrama, & Conger, 2008; South, Baumer, & Lutz, 2003; Taylor et al., 2010). A second implication of family SES is the degree to which it relates to family stress, instability, and neighborhood integration. Low-SES families tend to have a host of risk factors associated with elevated levels of family stress and poorer community integration (Albrecht & Albrecht, 2011; Melby et al., 2008; Ou, 2005; South et al., 2003; Taylor et al., 2010); risk factors may include frequent residential transitions, having young parents, or living in a single or unmarried household, all of which are related to lower educational attainment.

Family process factors also play a valuable role in children's educational attainment. Parents who have overly negative interactions with their children or who have personal problems that undermine effective parenting (e.g., couple issues) can impede their child's persistence in school (Dubow et al., 2009; King et al., 2006). Conversely, children whose parents are involved in their education, have a supportive parenting style, or hold high expectations for their educational attainment tend to stay in school longer (Ou, 2005; Pettit, Yu, Dodge, & Bates, 2009; Taylor et al., 2010). Robertson and Reynolds (2010) looked at the global influence of favorable family context by assigning students to clusters based on measures of demographic variables (e.g., mother age and education, number of adults living in the home, parental employment, subsidized meals) and of parenting (e.g., child maltreatment, parental involvement, parental expectations). Four clusters were found to be internally consistent in terms of human capital resources (based on demographic data) and family functioning. As predicted, children belonging to clusters that had higher levels of resources and high-quality parenting reached higher levels of educational attainment.

Numerous student characteristics have also been evaluated as predictors of future educational attainment, and they can be classified as risk or compensatory factors. Risk factors include predictors of poor academic adjustment, which can precipitate dropout or discourage involvement in higher education. Youth externalizing problems, especially when documented in childhood or early adolescence, have often been identified as predictors of lower educational attainment (King et al., 2006; McLeod & Kaiser, 2004; Pettit et al., 2009). Substance use later in adolescence also

has been consistently linked with poorer school persistence (Chatterji, 2006; Hardy et al., 1997; King et al., 2006; Ryan, 2010).

Compensatory factors that help facilitate progression through the education system have also been identified. They include students' educational aspiration and academic success (often assessed using grade-point average [GPA], standardized test scores, inclusion on the honor roll, avoidance of grade retention), which are strong and reliable predictors of educational attainment (Albrecht & Albrecht, 2011; Ganzach, 2000; Hardy et al., 1997; King et al., 2006; Marjoribanks, 2005; Mello, 2008; Ou, 2005; Pettit et al., 2009; South et al., 2003). Cognitive functioning, such as childhood IQ or general cognitive ability in early adulthood (Dubow et al., 2009; Kristensen et al., 2009), and positive psychological dispositions, including positive academic self-concept, academic engagement, future orientation, and positive temperamental dispositions (Beal & Crockett, 2010; Hampson, Goldberg, Vogt, & Dubanoski, 2007; Marsh & O'Mara, 2008; Melby et al., 2008), are also indicative of future educational attainment.

The extensive literature describing established risk and compensatory factors for educational attainment makes it possible to identify with considerable confidence students who are at high risk for leaving school before they obtain an adequate level of educational training. Because so many of these factors are difficult to alter, it is essential to identify student or parent characteristics that are amenable to change so that interventions can be developed to effectively bolster student retention, reduce dropout, and ultimately promote educational attainment (Rumberger, 1987). In an effort to help determine new predictors that have stronger implications for intervention research, we aimed in this study at testing effortful control as a predictor of educational attainment by age 23.

## Effortful Control

Effortful control is an aspect of temperament that reflects self-regulatory skill. Effortful control involves the ability to inhibit impulses and prevent disruptive behaviors (inhibitory control), to focus and maintain attention despite distractions (attention control), and to initiate and complete tasks that have long-term value, even when they are unpleasant (activation control; Rothbart & Bates, 1998).

Effortful control is heritable and shows moderate stability over time, but its development is also shaped by experience (Eisenberg et al., 2005; Goldsmith, Buss, & Lemery, 1997). Experimental studies have shown that aspects of effortful control can be improved in children, adolescents, and adults by a range of interventions, including mindfulness training (Sahdra et al., 2011; Tang et al., 2007), self-control exercises (Muraven, 2010), parent training (Somech & Elizur, 2012; Stormshak, Fosco, & Dishion, 2010), and school-based interventions (Diamond, Barnett, Thomas, & Munro, 2007; Raver et al., 2011).

A growing literature reveals that effortful control predicts academic success in children and adolescents, even after controlling for prior academic performance or general cognitive ability (Allan & Lonigan, 2011; Blair & Razza, 2007; Checa, Rodriguez-Bailón, & Rueda, 2008; Checa & Rueda, 2011; Valiente, Lemery-Chalfant, & Swanson, 2010; Valiente, Lemery-Chalfant, Swanson, & Reiser, 2008; Zhou, Main, & Wang, 2010). Posner and Rothbart (2007) have proposed that understanding the neurocognitive features of effortful control, its malleability, and its role in the growth



of competence in children is perhaps the most important agenda item for future research in education sciences. In fact, Posner and Rothbart propose that we should consider educating the human brain as much as teaching traditional content domains, such as reading, writing, and math. They contend that developing the neurocognitive skill of effortful control will benefit growth in general cognitive competence as much as in domain-specific skills. Although this idea is intriguing, relatively little research has examined it in general, let alone specific to adolescence and young adulthood. This omission is noteworthy in that adolescence is a turning point for many youths, at which time some disengage from academics and others persist into higher levels of educational attainment.

In our study, we extended findings about effortful control and academic success in childhood and adolescence by examining the relationship between effortful control and educational attainment in young adulthood. Effortful control may play a particularly important role in the pursuit and successful completion of postsecondary education. In comparison with earlier years of schooling, postsecondary education has unique qualities that make self-regulation especially important. Not only is postsecondary education voluntary, it also occurs within the developmental context of increasing freedom and responsibilities (Arnett, 2000). It requires that students manage the demands related to completion of their coursework and degree programs (time management, course selection, completion of long-term projects) in a context that provides less support and structure than is common in earlier levels of education. In addition, students are faced with the challenges of balancing the demands of their education with an expanding array of competing options and responsibilities that arise in emerging adulthood. Thus, it is expected that higher levels of effortful control will promote the planfulness that is involved in choosing to pursue higher education and the self-management that is required to successfully complete a degree. Consistent with this perspective, evidence is emerging that links school persistence and aspects of effortful control. For example, a recent study by Andersson and Bergman (2011) revealed that task persistence at age 13 was a statistically significant, albeit modest, predictor of educational attainment 30 years later. In addition, Wolfe and Johnson (1995) found that in predicting college GPA, self-discipline outperformed SAT standardized assessment scores. Although this preliminary research is promising, an important research goal is to determine whether effortful control predicts educational attainment.

### This Study

The aim of this study was to evaluate the role of effortful control in the progression toward higher levels of educational attainment in early adulthood. Because of policy and intervention implications of this study, we controlled for many of the family and individual variables that have historically predicted educational attainment so that we could conduct a more stringent test of the unique contribution of effortful control to educational attainment. Specifically, we controlled for key family processes, such as relationship quality and effective parenting practices, adolescent problem behavior during middle school, adolescent substance use and GPA during high school, and sociodemographic factors (family SES and parental education). We hypothesized that effortful control would be

a significant predictor of educational attainment, above and beyond established predictors.

Effortful control was assessed using parent, teacher, and adolescent self-report methods to create a multi-informant latent construct to ensure strong measurement of this focal construct in our study. Furthermore, we used a 12-year longitudinal design to represent the hypothesized sequence of action of different predictors and to avoid the inflated correlations that occur when predictors and outcomes are measured simultaneously. A secondary goal of this study was to verify whether our prediction model could generalize to students of both genders and to students of various ethnic groups.

To achieve these goals, we used structural equation modeling (SEM) to test the model presented in Figure 1. The hypothesized sequence of action of various predictors reflects the sensitive periods identified in the studies cited earlier in this article in relation to family situation, early adolescence problem behavior, substance use, and school adjustment as predictors of educational attainment. Positive family involvement and problem behavior are hypothesized to play an important role in early adolescence and to predict more proximal predictors of educational attainment, namely, substance use, high school cumulative GPA (CGPA), and effortful control in late adolescence. The possibility that early predictors are residually related to educational attainment about 10 years later is indicated by direct paths from early adolescence predictors to the outcome measure. To keep Figure 1 simple, we did not depict residual correlations among predictors measured during the same developmental period, but they were included in the statistical model (i.e., problem behavior was correlated with positive family involvement; substance use, high school CGPA, and effortful control were intercorrelated; and family SES and parental education were correlated with each other and with the five other predictors in the model). Our primary analyses were conducted on the entire sample, and we tested the generalizability of our findings to various subgroups by using multiple-group analyses.

## Method

### Participants

Participants were 997 adolescents and their families from the Project Alliance 1 study recruited in Grade 6 from three public middle schools in an ethnically diverse metropolitan community in the northwestern United States. Parents of all Grade 6 students in two cohorts (years 1996 and 1998) were approached for participation, and 90% consented. The participating sample included 526 males (52.8%) and 471 females (47.2%). By youth self-report, the sample comprised 423 European Americans (42.4%), 291 African Americans (29.2%), 68 Latinos (6.8%), 52 Asian Americans (5.2%), and 164 (16.4%) youths of other ethnicities, including mixed ethnicity. Parent reports collected when the adolescents were 16 years old revealed that 39.6% of participants lived with both genetic parents, 43.8% lived with their biological mother, 6.7% lived with their biological father, and 10.0% lived in other family configurations. The median range of gross annual household income was \$30,000–\$39,999, with 25.3% of households earning less than \$20,000 per year and 12.7% earning more than \$90,000.

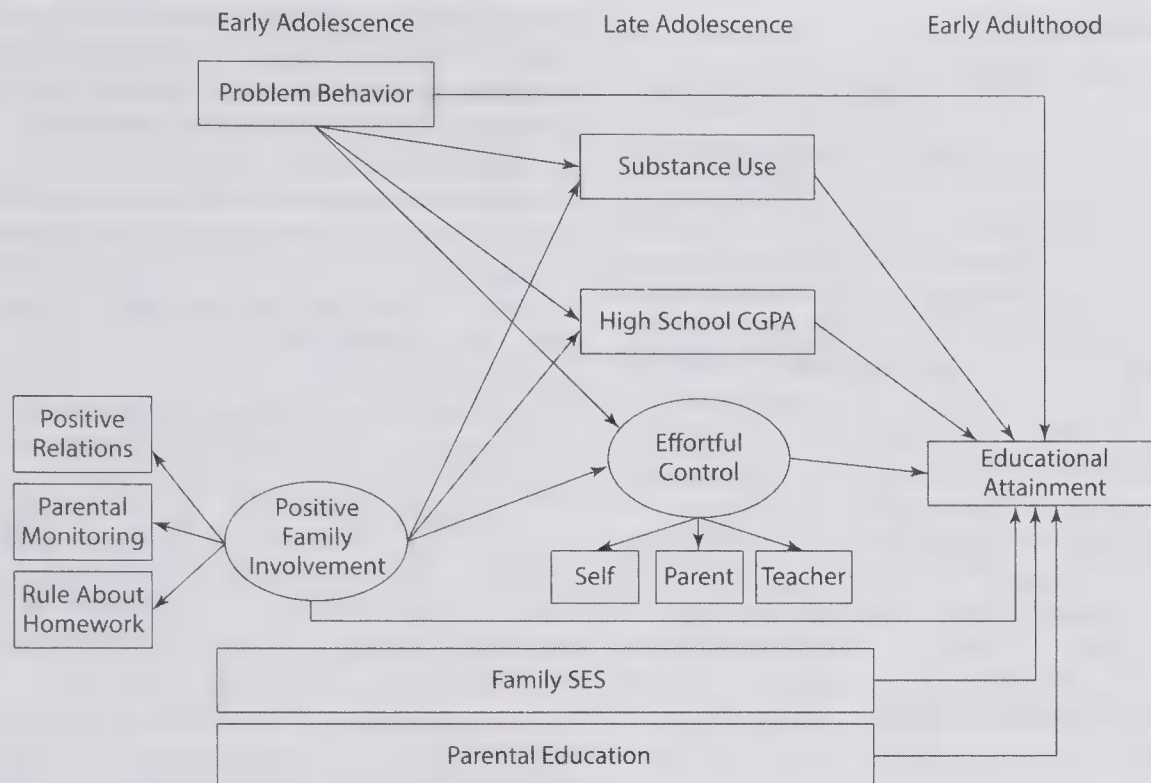


Figure 1. Full model. Correlations among predictors measured within the same developmental period were also included in the model, even if they are not depicted here. Correlations between family socioeconomic status (SES) and parental education, and between these variables and the five other predictors, were included. CGPA = cumulative grade-point average.

Because most participants remained in the same middle school from Grade 6 through Grade 8, and because data collection took place in the school setting, a high rate of retention was maintained across the first three time points. Most participants were streamed into a few local high schools whose principals agreed to help us track participants, which greatly facilitated data collection in Grades 9 and 11. These procedures, however, were not sufficient for participants who stopped attending the schools involved in our study and were not useful after participants graduated from high school. Additional procedures were therefore put in place; namely, at each time point, participants were asked to fill out a form with their current contact information (mailing address, phone numbers) and to provide the contact information of other people (e.g., friends, family members) who could help us find them if they had moved before the next time point of our data collection. Participants were also paid \$5 for sending us their new contact information when they moved. Under those circumstances, questionnaires that were usually filled out in school could be filled out at home and mailed back to us. Together, these longitudinal retention procedures were very effective, with approximately 80% of youths being retained across the study span.

One half of the study sample was randomly assigned to a multilevel family-centered ECOfit (Dishion & Kavanagh, 2003; Dishion & Stormshak, 2007), which aimed at preventing substance use and problem behavior in adolescents. Intent-to-treat analyses revealed positive intervention effects in relation to substance use (Connell, Dishion, & Deater-Deckard, 2006), antisocial behavior (Van Ryzin & Dishion, 2012), and the probability of police arrest (Connell, Klostermann, & Dishion, 2012). In addition, using complier average causal effect analyses to assess the impact of fami-

lies' engagement in the selected level of this intervention (the Family Check-Up), we found significant intervention effects on substance use, problem behavior, school grades, and attendance during middle and high school (Connell, Dishion, Yasui, & Kavanagh, 2007; Stormshak, Connell, & Dishion, 2009; Véronneau, Dishion, & Connell, 2013). Because improving educational attainment was not a goal of this program and because traditional intent-to-treat effects were not found for academic outcomes in middle and high school, we did not expect major differences in the covariance matrices of the intervention and control groups based on the variables of interest in this study. To verify this assumption, we used participants' raw data while testing for equivalence of the unconstrained covariance matrices for the treatment and control groups and found good model fit for most, but not all indices,  $\chi^2(76) = 110.02, p < .01$ , root-mean-square error of approximation (RMSEA) = .03, comparative fit index (CFI) = .98, Tucker-Lewis Index (TLI) = .97. The chi-square test suggests that we should reject the null hypothesis, stating that the treatment and control groups have equivalent covariance matrices; in contrast, all other fit indices suggest that constraining the covariance matrices of the two groups yields a well-fitting model, with both CFI and TLI > .95 and RMSEA < .06 (Hu & Bentler, 1999). Because the chi-square test may be overly sensitive to trivial group differences when large sample sizes are used (as is the case in this study), we prioritized the other fit indices and concluded that the two groups did not differ with regard to the covariance of our study's variables. Therefore, data from the two groups were pooled in this study's analyses.



## Assessment Procedures

School-based self-report assessments of problem behavior and family involvement were collected from students in Grades 6 through 8 using an adaptation of a survey instrument developed and reported by scientists at the Oregon Research Institute for the Community Action for Successful Youth project (Metzler, Biglan, Ary, & Li, 1998). In Grade 11, a larger assessment protocol was conducted that included additional student self-report surveys, teacher ratings that were administered in the high school setting, and parent report questionnaires that were completed at home and mailed to our research office. This Grade 11 assessment was the final school-based assessment. After high school, subsequent assessments were conducted when participants were approximately 19 years old and again when they were approximately 23 years old. The age 19 assessment was limited in scope and did not pertain to the current study. However, age 23 questionnaires captured constructs of interest to our study. At this wave, questionnaires were sent directly to participants' homes and were returned to our research office by mail. All respondents were assured of the confidentiality of their responses. Participants, parents, and teachers were compensated for their participation.

## Measures

**Family SES.** SES was measured by parent report of their employment status, income, housing status, and financial aid to the family. For employment status, we used the highest score based on reports from both primary caregivers when participants were from two-parent families (*full time or self-employed* [coded 4]; *part time* [3]; *seasonal* [2]; *disabled, unemployed, temporary layoff, home-maker, retired, or student* [1]). One global score was used for each of the other indicators: family housing (*own your home* [coded 5], *rent your home* [4], *motel/temporary* [3], *live with a friend or live with a relative* [2], and *emergency shelter or homeless* [1]); household income (*\$90K or more* [coded 7], *between \$70K and \$90K* [6], *between \$50K and \$70K* [5], *between \$30K and \$50K* [4], *between \$20K and \$30K* [3], *between \$10K and \$20K* [2], and *less than \$10K* [1]); and financial aid (sum of dichotomous indicators of whether the family received food stamps, Aid to Families with Dependent Children, other welfare, medical assistance, and Social Security death benefits, reverse coded). These variables were standardized and averaged ( $\alpha = .75$ ). In this study, SES information was not collected from youths because of concern that it would potentially be unreliable information. The Grade 11 data collection was the first time point when all parents were surveyed, and thus this is the earliest wave of SES data for the overall sample. SES was not assigned to a specific developmental period in the model and was treated as a fixed variable that other predictors from any time point could be correlated with.

**Parental education.** At the Grade 11 assessment, caregivers reported on the highest level of education that they themselves had achieved: *graduate degree or college degree* (coded 5), *junior college or partial college* (4), *high school graduate* (3), *partial high school or junior high completed* (2), and *7th grade or less or no formal schooling* (1). When data were provided for two primary caregivers, we used the highest of the two scores.

**Positive family involvement.** This latent variable was measured from a combination of three youth report indicators. For each of these three indicators, an average score based on data collected

in Grades 6, 7, and 8 was computed as a reliable index for the entire middle-school period. The first indicator, positive family relations, was based on a six-item scale that included statements such as "I really enjoyed being with my parents," "My parents trusted my judgment," "Family members backed each other up." Each item was scored on a scale ranging from 1 (*never true*) to 5 (*always true*) within the past month, and a mean score was computed from the six items ( $\alpha$ s for Grades 6 through 8 ranged from .89 to .90). The second indicator, parental monitoring, was based on a five-item scale that asked the youths how often their parents knew what they were doing away from home; where they were after school; what their plans were for the next day; and what were their interests, activities, and whereabouts. Each item was scored on a scale ranging from 1 (*never or almost never*) to 5 (*always to almost always*), and a mean score was created on the basis of all five items ( $\alpha$ s for Grades 6 through 8 ranged from .85 to .87). The third indicator, homework rule, included one item that reflected whether parents had a rule about the child doing homework every day. The item was scored on a scale ranging from 1 (*don't have a rule or expectation*) to 4 (*have a clear rule*).

**Problem behavior.** Problem behavior was measured using a nine-item self-report scale administered in Grades 6, 7, and 8. The variable was created from an average score based on data collected at all three time points to create a reliable measure for the entire middle-school period. Sample items include "Stayed out all night without parents' permission," "Intentionally hit or threatened to hit someone at school," and "Stole or tried to steal things worth more than \$5." Each item was rated on a scale ranging from 1 (*never*) to 6 (*more than 20 times*), and the reference period was during the past month ( $\alpha$ s at Grades 6 through 8 ranged from .77 to .84).

**Effortful control.** The three indicators of the effortful control construct were administered in Grade 11: parent report, self-report, and teacher report. Parent and child reports were based on the Effortful Control scale from the short form of the Early Adolescent Temperament Questionnaire-Revised (EATQ-R; Ellis & Rothbart, 2005). The EATQ-R Effortful Control scale consists of 16 items that assess activation control (the capacity to perform an action when there is a strong tendency to avoid it; e.g., "If I have a hard assignment to do, I get started right away"), attention (the capacity to focus attention as well as shift attention when desired, e.g., "It is really easy for me to really concentrate on homework problems"), and inhibitory control (the capacity to plan and to suppress inappropriate responses, e.g., "I can stick with my plans and goals"). Each item was scored on a scale ranging from 1 (*almost always untrue*) to 5 (*almost always true*), with higher scores indicating greater effortful control.

Previous work by Ellis and Rothbart (2001) reports evidence of the validity of the Effortful Control scale for a sample of adolescents ranging in age from 10 to 16. Their study demonstrated adequate internal consistency ( $\alpha = .80$  for the self-report,  $\alpha = .87$  for the parent report) and acceptable convergence ( $r = .50$ ) between adolescent and parent report (Ellis, 2002). The self- and parent report versions include essentially the same items, with the pronouns changed appropriately. For the parent reports, participants' mothers, fathers, and other guardians could all complete the Effortful Control scale. When multiple caregivers responded, those answers were averaged into one parent report score. Internal consistency for the 16-item Effortful Control scale was .63 for youths, .77 for mothers, and .82 for fathers.



The third indicator, teacher report of effortful control, consisted of five items with content similar to that of the EATQ-R Effortful Control scale (e.g., “thinks ahead of time about the consequences of actions,” “plans ahead before acting,” “pays attention to what he or she is doing,” “works toward goals,” and “sticks to what he or she is doing until it is finished, even with unpleasant tasks”). Teachers used a 5-point rating scale to describe how frequently each participant engaged in these behaviors. The internal consistency of the teacher report scale was  $\alpha = .94$ .

**CGPA.** Students’ academic records were gathered from the schools from Grade 9 through 11. If a participant moved to another school, we sought academic records from the new school as well. GPA was measured on a scale ranging from 0 to 4, with higher scores reflecting better grades (F = 0, D = 1, C = 2, B = 3, A = 4). GPA was obtained at the end of each school year as the average grade across participants’ academic courses for that year. For youths who attended multiple schools during an academic year, an adjusted GPA was computed as the average of the available GPAs, weighted to reflect the proportion of the school year they represented. Our analyses used a CGPA measure computed as the average of all yearly GPA data available for Grades 9 through 11. For the cohort of participants who were originally enrolled in 1998 (about half the participants), CGPA in Grade 11 was unavailable because of a change in the school district’s record-keeping system. Other students had missing GPA data because of school dropout or because they attended schools that were unable to provide official academic records. As a result, 47% of participants had a CGPA measure based on all 3 years of high school; 30% had a CGPA measure based on Grades 9 and 10, and 12% had a CGPA based on Grade 9 only, resulting in 89% of participants with valid GPA data for the main analyses. Correlations between CGPA and yearly GPA were .80 for Grade 11, .93 for Grade 10, and .93 for Grade 9 (all  $ps < .001$ ).

**Substance use.** Participants completed a survey in Grade 11 that enabled us to measure the extent of their substance use. Participants reported on their use of tobacco, alcohol, marijuana, and other drugs, and an average score for substance use was created. Participants were asked to report their frequency of use during the past 3 months for each substance, on a scale ranging from 0 (*never*) to 7 (*2 or 3 times a day or more*). “Other drugs” was defined for the participants as any of the following substances: heroin, morphine, cocaine or crack, speed or meth, ecstasy, angel dust or PCP, acid or LSD, mushrooms, gasoline, glue, other inhalants, and prescription medications for recreational use.

**Ethnicity.** Although various ethnic groups were represented in this sample, only the two largest groups (European American and African American) could be used for ethnic comparison purposes, and we used youth report of their ethnicity.

**Educational attainment.** Participants reported on the highest level of education they had completed as of the age 23 assessment. This information was coded on a 4-point scale: *less than high school* (coded 1), *high school/GED* (2), *trade school/some college/specialized training/2-year college degree* (3), or *4-year college or graduate degree* (4). This measure was treated as an ordered categorical variable in the primary analyses.

## Results

### Preliminary Analyses

**Missing data.** For the variables included in our study, the mean percentage of missing data was 14% (range = 0%–33%). Little’s missing completely at random (MCAR) test was significant,  $\chi^2(361) = 505.54, p < .001$ , indicating that the data were not MCAR. We explored patterns of missingness based on the amount of missing data for different subgroups of participants by counting the number of variables for which there was a missing value for each participant. Then, we examined correlations between the total number of missing values for each participant and their scores on other measured (i.e., nonmissing) variables.

Missing data were more common among male participants and among participants with lower educational attainment, lower CGPA, lower SES, lower parental education, lower parent-reported effortful control, less parental monitoring, and more substance use ( $rs = .08$ – $.18, ps < .05$ ). Missingness differed significantly across ethnic groups,  $F(2) = 4.66, p < .01$ . When comparing European Americans, African Americans, and other minority groups combined, a post hoc Scheffé test revealed that participants from other minority groups had more missing data than did European American participants (mean difference = 1.21;  $p < .05$ ).

Covariance coverage was moderate to high, ranging from .59 to 1.00. Full information maximum likelihood (FIML) was used within Mplus 7.0 to estimate parameters on the basis of all available information from each participant. Consequently, participants with occasional missing data were retained in the analyses. FIML has been shown to be very efficient when analyzing data from samples with moderate levels of missing values, and it is adequate even when data are not MCAR, as long as the predictors of missingness are included in the model (Widaman, 2006).

**Descriptive statistics and correlations.** Means, standard deviations, and correlations among all measured variables are presented in Table 1, along with the number of participants who provided valid data on each measure, and skewness and kurtosis values. Early problem behavior had a skew value greater than 2.0 and a kurtosis value greater than 8.0 (cutoffs provided by Kline, 2005) and was thus square-root transformed. The transformed variable did not have significant skew or kurtosis and was used in all subsequent analyses. All other variables were approximately normally distributed (skew  $< 2.0$  and kurtosis  $< 8.0$ ). As expected, educational attainment had a strong positive correlation with CGPA; a moderate positive correlation with family SES, parental education, and effortful control according to the teacher; and a weaker but significant positive correlation with positive family relations, parental monitoring, homework rule, and both self-report and parent report of effortful control. Educational attainment had a weak but significant negative correlation with early adolescence problem behavior and late adolescence substance use. CGPA, measures of positive family involvement, parental education, and SES were negatively correlated with measures of problem behavior and substance use.

**Group differences.** Gender and ethnicity differences in all observed variables were examined with a series of one-way analyses of variance. Females had higher educational attainment, higher CGPAs, higher ratings on caregiver and teacher reports (but



Table 1  
Descriptive Statistics and Bivariate Correlations

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1. Educational attainment	—											
2. Family SES	.32***	—										
3. Parental education	.45***	.42***	—									
4. Early problem behavior	-.28***	-.19***	-.23***	—								
5. Positive family involvement: positive relations	.11**	.01	.01	-.29***	—							
6. Positive family involvement: parental monitoring	.22***	.12***	.19***	-.53***	.51***	—						
7. Positive family involvement: homework rule	.11**	-.001	.05	-.25***	.30***	.33***	—					
8. Substance use	-.15***	.02	.02	.27***	-.18***	-.19***	-.16***	—				
9. CGPA	.60***	.39***	.40***	-.30***	.12**	.26***	.11**	-.19***	—			
10. Effortful control–self	.12***	-.09*	-.004	-.15***	.21***	.17***	.12**	-.16***	.12***	—		
11. Effortful control–parent	.27***	.06	.03	-.18***	.18***	.19***	.06	-.21***	.37***	.34***	—	
12. Effortful control–teacher	.45***	.25***	.22***	-.23***	.09*	.20***	.06	-.17***	.57***	.20***	.41***	—
Mean	2.65	.01	3.94	1.40	3.47	3.97	3.38	.73	2.17	3.35	3.30	3.72
SD	.90	.75	.97	.48	.86	.80	.62	1.13	1.08	.48	.54	.79
n	855	726	706	997	997	997	995	792	884	792	684	666
Skew	-.15	-1.04	-.82	2.43	-.34	-.94	-1.17	1.72	-.16	.26	-.20	-.18
Kurtosis	-.74	.41	.31	8.08	-.46	.57	1.37	2.17	-.86	.34	-.03	-.56

Note. SES = socioeconomic status; CGPA = cumulative grade-point average.

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

not self-reports) of effortful control, less early adolescence problem behavior, and more early adolescence parental monitoring than did males (all  $F$ s  $> 10.0$ ,  $p$ s  $\leq .001$ ).

Ethnic differences were obtained for all measures except caregiver-reported effortful control. African American participants had higher self-reported effortful control but lower teacher-reported effortful control relative to Caucasian participants. CGPA, parental education, SES, and parental monitoring were lower for African American participants, and homework rule and positive family involvement were higher for African American participants. African American participants reported more problem behavior in early adolescence but less substance use in late adolescence, relative to Caucasian participants (all  $F$ s  $> 4.45$ ,  $p$ s  $< .05$ ).

## Primary Analyses

Hypothesis testing proceeded in two steps: evaluation of the hypothesized model (see Figure 1) and examination of group differences (gender and ethnicity) in model fit. We evaluated the fit of the hypothesized model to the data using Mplus 7.0. SEMs were run using the mean- and variance-adjusted weighted least square estimator because the outcome variable (educational attainment) was ordered categorical. Therefore, parameter estimates for the predictors of educational attainment can be interpreted as probit regression coefficients. Residual errors were allowed to correlate for latent-variable indicators with shared measures (i.e., child- and parent-reported effortful control, both of which used the EATQ-R questionnaire) and/or shared reporters (i.e., child-reported indicators of effortful control and family involvement). The model was deemed to have adequate fit if the CFI was  $> .95$ , and the RMSEA was  $< .06$  (Hu & Bentler, 1999). Good model fit

is usually indicated by nonsignificant chi-square values, but because of the large size of our sample, this index of fit may be overly conservative (Schermelleh-Engel, Moosbrugger, & Müller, 2003). In this situation, it is common practice to give priority to the other fit indices in model fit evaluation.

To examine group differences, we ran a series of multiple-group analyses (for gender and ethnicity) and compared model fit for unconstrained models (all regression and correlation coefficients free to vary across groups) and constrained models (coefficients constrained to be equal across groups). Because of the large sample size, we used change ( $\Delta$ ) in CFI to test for the significance of differences in fit. Fit was considered to be significantly different if the change in CFI was .01 or greater (Cheung & Rensvold, 2002).

The hypothesized model provided a good fit to the data,  $\chi^2(29) = 116.18$ ,  $p < .001$ , CFI = .96, RMSEA = .06. Standardized coefficients for regression paths and factor loadings are presented in Figure 2. There were three significant predictors of educational attainment: adolescent effortful control ( $\beta = .33$ ,  $SE = .09$ ), parental education ( $\beta = .29$ ,  $SE = .04$ ), and high school CGPA ( $\beta = .26$ ,  $SE = .08$ ). All three predictors had effect sizes in, or very close to, the moderate range. We built a 95% confidence interval around these coefficients to test the null hypothesis that these predictors were of equal strength, and we were unable to reject it. This suggests that adolescents with higher levels of effortful control at age 17 had higher levels of educational attainment by age 23, and the unique relation of effortful control with future educational attainment is comparable in strength to that of other well-established predictors. Other control variables used in this model were not statistically significant predictors of educational attainment, including family SES, problem behavior, and

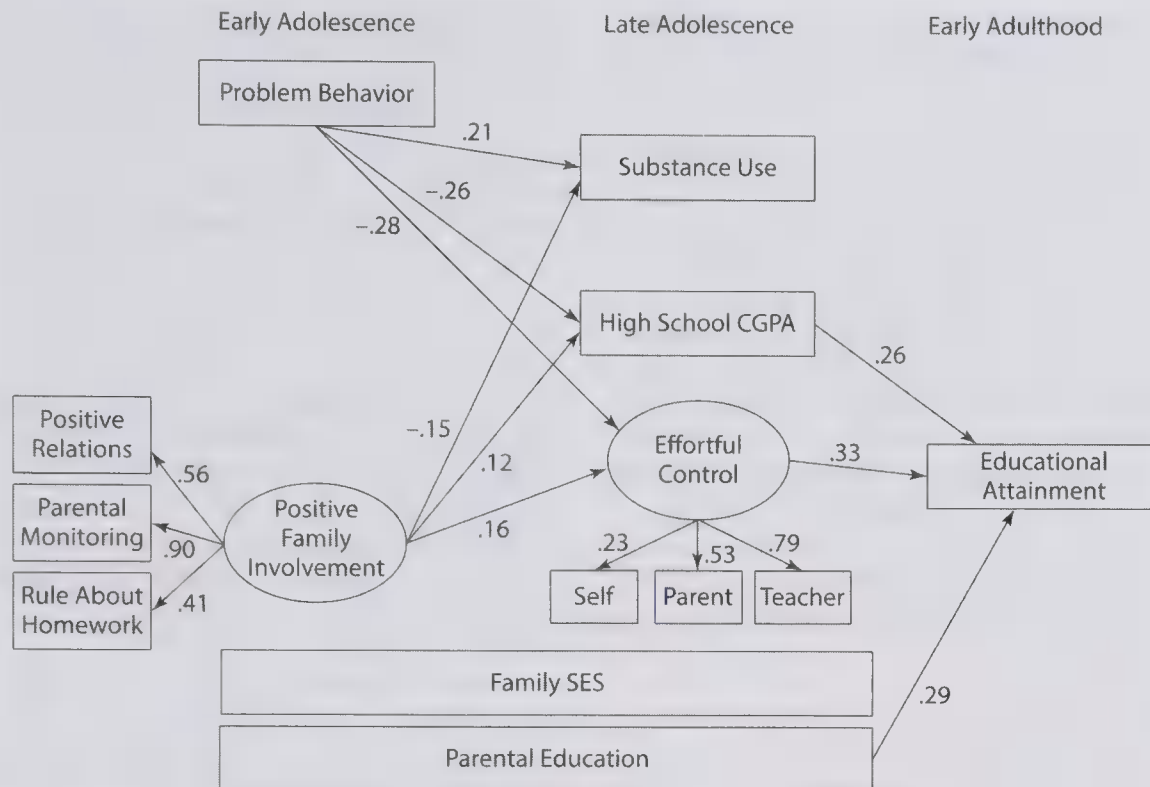


Figure 2. Model results (regression paths and factor loadings). Coefficients are standardized. All solid paths are significant at  $p < .05$  or smaller. Other regression paths mentioned in Figure 1 that are not depicted here were included in the structural equation modeling analyses but were not significant. SES = socioeconomic status; CGPA = cumulative grade-point average.

family involvement in early adolescence. Similarly, late adolescence substance use was not associated with educational attainment. These nonsignificant paths are omitted from Figure 2 for parsimony, but they were still present in the statistical model. Correlations that were modeled between residual errors because of shared measures or reporters were positive and significant ( $r_s = .10-.28, p_s < .01$ ). The estimated correlation matrix for the latent variables in the model is presented in Table 2. Model-estimated residual correlations among variables that were measured within the same developmental period were identical to those reported in Table 2, except for the following: Family SES correlated significantly with substance use ( $r = .08, p < .05$ ), CGPA ( $r = .35, p < .001$ ), and effortful control ( $r = .15, p < .01$ ); parental education correlated significantly with substance use ( $r = .10, p < .01$ ), CGPA ( $r = .35, p < .001$ ), and effortful control ( $r = .11, p < .05$ );

substance use correlated significantly with CGPA ( $r = -.10, p < .05$ ) and effortful control ( $r = -.22, p < .001$ ); and CGPA correlated significantly with effortful control ( $r = .67, p < .001$ ).

Tests of indirect effects were performed using confidence intervals based on the bias-corrected bootstrap method (MacKinnon, Lockwood, & Williams, 2004) to verify whether the late adolescence predictors—effortful control and academic achievement—could explain the relation between early adolescence predictors (family involvement and problem behavior) and educational attainment. Results revealed that effortful control was a significant mediator for none of the early adolescence predictors. CGPA was a marginally significant mediator of the relationship between early adolescence family involvement and educational attainment, with a 90% confidence interval for the  $\beta$  value ranging from .003 to .076 (point estimate = .032). Furthermore, CGPA was a signifi-

Table 2  
Estimated Correlation Matrix for the Latent Variables

Variable	1	2	3	4	5	6	7	8
1. Educational attainment	—							
2. Family SES	.34***	—						
3. Parental education	.48***	.42***	—					
4. Early problem behavior	-.32***	-.21***	-.25***	—				
5. Positive family involvement	.25***	.09*	.16***	-.59***	—			
6. Substance use	-.17***	.02	.02	.29***	-.27***	—		
7. CGPA	.64***	.40***	.41***	-.33***	.27***	-.19***	—	
8. Effortful control	.59***	.21***	.20***	-.37***	.32***	-.32***	.71***	—

Note. SES = socioeconomic status; CGPA = cumulative grade-point average.

\*  $p < .05$ . \*\*\*  $p < .001$ .



cant mediator of the relation between early adolescence problem behavior and educational attainment, with a 99% confidence interval for the  $\beta$  value ranging from  $-.131$  to  $-.005$  (point estimate =  $-.068$ ).

Group invariance tests were conducted to determine whether differences in model fit were evident across groups, which would suggest moderation effects based on gender or ethnicity. Tests for group differences in model fit revealed no significant differences between constrained and unconstrained models for gender ( $\Delta\text{CFI} = .002$ ). The pattern of results obtained from a pooled within-group covariance matrix was identical to the one presented in Figure 2. In line with preliminary analyses, multiple-group analyses comparing ethnic groups (Caucasian vs. African American) revealed that constraints imposed on mean levels of several variables had to be released. These included family SES, teacher rating of self-regulation, and parental monitoring. The constraint on the residual (unexplained) variance for educational attainment was also relaxed. This new model did not differ significantly from the unconstrained model.

## Discussion

The main objective of this study was to test whether adolescents' effortful control is a significant predictor of their educational attainment in early adulthood, above and beyond established academic, familial, behavioral, and demographic factors. The significant relationship between effortful control and educational attainment supported our hypothesis, and follow-up analyses revealed that the final model applied to both genders and was generalizable across European American and African American participants.

### Effortful Control as a Predictor of Educational Attainment

*Effortful control* is defined as a temperament-based individual characteristic that reflects self-regulatory skill, manifested by the ability to inhibit impulses and disruptive behaviors (inhibitory control), to focus and maintain attention in spite of distractions (attention control), and to initiate and complete tasks that have long-term value (activation control; Rothbart, Ellis, & Posner, 2011). In this study, we tested whether effortful control was related to educational attainment after accounting for other well-documented predictors. After controlling for other factors, effortful control was directly associated with educational attainment. Moreover, our findings indicate that effortful control is as important as parental education and high school academic achievement for predicting educational attainment in early adulthood.

Several mechanisms could explain the relationship between effortful control and educational attainment and should be investigated in future studies. One possibility is that as students progress through the late high school years and postsecondary education, they must increasingly rely on their own volitional resources as parents and teachers step out of their supervisory responsibilities to encourage students' autonomous academic development. Adequate levels of effortful control may support the planfulness and self-management needed to successfully complete a postsecondary degree. In addition to increased demands on students' autonomy and planning skills, the changing nature and context of the school-

work required of them can also represent a significant change in their academic life. Being able to adapt their work habits accordingly (e.g., creating study groups; starting to work on assignments many weeks before the deadline) and to maintain these new behaviors over the long term instead of persisting with or going back to old habits that may not be adaptive in this new context could be one way in which effortful control influences academic success and persistence.

Our findings are consistent with those presented in past studies that have explored other constructs related to effortful control as predictors of educational and professional success in adulthood (Andersson & Bergman, 2011; Wolfe & Johnson, 1995). This study builds on existing literature that underscores the importance of parental education and youth academic success as key predictors of educational attainment. Beyond parental support and academic ability, adolescents' self-regulatory capacity inherent in effortful control makes a compelling argument for the importance of targeting effortful control in efforts to promote school persistence.

Our study findings also are consistent with those from past studies that have identified processes that can promote effortful control functioning (e.g., Fosco, Frank, Stormshak, & Dishion, 2013; Muraven, 2010; Stormshak et al., 2010). Although our study was not designed to test for predictors of effortful control, we did identify direct links between positive family involvement and problem behavior during early adolescence and later effortful control; however, we were unable to find significant indirect effects involving adolescent effortful control as a mediator of positive family involvement in the prediction of educational attainment. Nevertheless, the role of parenting in promoting effortful control is supported by other research, including a study by Bowers et al. (2011), showing that aspects of self-regulation closely related to effortful control tend to decrease during adolescence but can increase under conditions of good parental practices, as do GPA and school attendance.

### Early Adolescence Predictors

Previous work that had investigated the contribution of early adolescence predictors of educational attainment prompted us to expect a negative relationship between problem behavior and future levels of educational attainment, and a positive relationship between positive family involvement—including the quality of relationships, parental monitoring, and rules about doing homework—and educational attainment. However, in our model, the direct paths between these factors from early adolescence and educational attainment were not significant. Instead, our findings suggest that these relationships are mediated by more proximal factors, such as academic achievement, which was a moderately strong predictor of educational attainment. The indirect effects of problem behavior and of family involvement on educational attainment support the idea that these early predictors do matter and deserve attention from researchers and practitioners who seek to promote educational attainment in youths beginning at an early age.

Regarding the family-related predictors, it had already been established that warm but structuring parenting can facilitate academic achievement and discourage adolescents' substance use (Coombs & Landsverk, 1988; Leung, Lau, & Lam, 1998; Stein-



berg, Lamborn, Dornbusch, & Darling, 1992). Of particular interest to us, though, was the possibility that family involvement could help promote greater effortful control during adolescence, as suggested by recent studies (Bowers et al., 2011; Doan, Fuller-Rowell, & Evans, 2012). Without any earlier measurement of effortful control in the sample, it was not possible to verify whether this relationship simply reflected the enduring consequences of parent-child dynamics promoting effortful control early in childhood. Nevertheless, Stormshak, Fosco, and Dishion (2010) found that a parent-focused intervention was related to an improvement in their children's effortful control over time, which supports the view that parents can actively help their child develop higher levels of effortful control in middle school. Effortful control was, in turn, predictive of an increase in academic engagement in high school. This is a new and promising avenue for applied research in the domain of academic persistence. In fact, the role of family relationships may be particularly consequential, considering a study by Belsky and Beaver (2011) that suggested that genetic predispositions can make male adolescents particularly vulnerable to deficits in self-regulation when they are exposed to poor parenting practices.

The association between problem behavior in childhood and academic outcomes in adolescence has already been documented (Véronneau, Vitaro, Pedersen, & Tremblay, 2008), but the path from problem behavior to effortful control is of greater interest in this study. The absence of repeated measures of problem behavior and effortful control makes it difficult to settle with confidence on a specific direction of a possible causal effect. Numerous studies have linked lower levels of effortful control (or related self-regulatory skill) in early childhood to later development of problem behaviors (e.g., Eiden, Edwards, & Leonard, 2007; King, Lengua, & Monahan, 2013; Lengua, 2006; Robins, John, Caspi, Moffitt, & Stouthamer-Loeber, 1996). However, our study also supports the possibility that young adolescents who engage in antisocial activities may, as a result, be diverted from opportunities to practice and reinforce their ability to exert effortful control—for example, by being suspended from school.

### Late Adolescence Predictors

Previous research that had suggested that substance use in later adolescence and academic achievement in high school could help predict which students would reach higher levels of educational attainment motivated us to include these two predictors as concurrent control variables when testing for effortful control as a predictor of educational outcomes.

Although the nonsignificant role of substance use in this study contrasts with results from other studies that revealed a significant role with a similar outcome, several explanations for the discrepant results are possible. For example, Hardy et al. (1997) found that smoking cigarettes in adolescence is related to lower levels of education in adulthood, but their educational outcome distinguished only between students who obtained a high school diploma/graduate equivalency degree and those who did not. Furthermore, their study assessed inner-city children who had been born in the 1960s, in contrast with our participants who were from a wider range of demographic backgrounds and who had been born in the 1980s. Cohort effects or differences in demographic backgrounds could explain the divergence of results between their

study and ours. Ryan (2010) found a detrimental contribution of marijuana use, but again the sample of participants had been born much earlier (the late 1950s to early 1960s), and the control variables used in this study focused more heavily on sociodemographic characteristics than on family dynamics and students' academic achievement. A study by King et al. (2006) in which a more comparable set of control variables was used revealed, by using growth modeling techniques, a significant contribution of drug use (but not alcohol use) to the likelihood of attending college. This finding suggests that research focused on the specific contribution of substance use to educational attainment would benefit from sophisticated longitudinal modeling of such variables. Because effortful control has already been shown to reduce the risk of increases in tobacco and marijuana use over time (Piehler, Véronneau, & Dishion, 2012), it is interesting to note that the association between effortful control and educational attainment in our study was completely independent from substance use. In other words, it is unlikely that the link between effortful control and higher educational attainment can be explained merely by the capacity to refrain from using substances.

Consistent with past research, our study revealed that academic competence in high school as measured from school records of academic achievement (CGPA) was a significant predictor of educational attainment, and its influence was moderate in size, just like that of effortful control and of parental education. It should be noted that CGPA was highly correlated with effortful control ( $r = .71$  for the estimated bivariate correlation, and  $r = .67$  for the model estimated residual correlation). The strong correlation between academic achievement and effortful control is consistent with results from theoretical work and empirical work linking effortful control to academic performance (e.g., Allan & Lonigan, 2011; Checa et al., 2008; Posner & Rothbart, 2007; Valiente et al., 2010). Longitudinal studies with repeated measurements of both effortful control and academic achievement would help confirm the sequence of action of effortful control and academic achievement that predict educational attainment.

### Sociodemographic Factors

In line with past research, students' sociodemographic background played an important role in the prediction of educational attainment. We expected that families with higher income and more stable living conditions (e.g., owning or renting a house rather than living in a precarious housing situation) are in a better position to support their child through their high school studies and provide financial resources that facilitate access to higher education. Although a moderate correlation emerged between family SES and participants' educational attainment in preliminary bivariate analyses, this association was not significant in the overall model, when controlling for other predictors. In contrast, parent education was linked to children's educational attainment in the overall model, independent of family financial resources. This finding provided support for our decision to examine the influence of parent education and that of other SES indicators separately. Our study cannot speak to the mechanisms linking parent education to child educational attainment in this sample, but numerous plausible explanations have been identified by other studies, including parental involvement, parents' ability to understand and navigate the school system, parental expectations, and family



attitudes toward schooling (e.g., Martin, 2012; Pettit et al., 2009). Given that effortful control is partly heritable, the link between parents' and children's educational attainment might also reflect genetic predispositions for self-regulatory skills that support school success and persistence.

### Strengths and Limitations

This study possesses many strengths. First, our main predictor, effortful control, was based on a latent variable that included parent, teacher, and self-reports. Also, we were able to control for most of the established predictors of educational attainment, which strengthens our conclusions about the significant role of effortful control in predicting our outcome of interest. In addition, the longitudinal design made it possible to use predictors at important times of development from early to late adolescence and to assess educational attainment in early adulthood, when a good level of variance has emerged in this variable. The large number of participants helped us identify small effects and compare results across subgroups of participants (gender, ethnicity). It is noteworthy that the relationships between the many predictors in this model and educational outcomes were consistent across genders and ethnic groups (European American vs. African American). This suggests that concrete interventions based on the results from this study are likely to be relevant for most students. Further research that includes a larger number of students belonging to the smaller ethnic groups is needed, however, to verify whether our results generalize to them.

Some limitations in this study would be useful to consider in future work. Having access to earlier measurements of effortful control would have been very helpful to test its contribution to educational attainment from a process standpoint. For example, effortful control at an early age could affect educational attainment through its influence on academic achievement, family relationships, or other mediators. Repeated measures of effortful control could even help determine whether it can be increased through environmental influences or intervention programs. To help explore those possibilities, a more recent study by our research group (Project Alliance 2; Stormshak et al., 2010) included several measures of effortful control completed during the adolescent years. Another limitation is that this study had no measure that allowed us to control for students' educational aspiration, which has been shown in past research to be a significant predictor of educational attainment (e.g., Dubow et al., 2009; Marjoribanks, 2005; South et al., 2003). In addition, the longitudinal nature of the study led to some missing-data issues. In general, missing data was more common among males and among lower functioning adolescents (lower CGPA, lower effortful control, more substance use) and parents (lower SES, lower parental education, less parental monitoring). These patterns might limit the generalizability of our results and suggest that lower functioning participants might have had more difficulty responding to the questionnaires, possibly because of lower reading abilities or because of additional stressful life events that may leave them less time or less availability to answer a questionnaire. Still, by using FIML to manage missing values, we are confident that our results are less biased than those we would have obtained using other popular strategies (e.g., list-wise deletion, mean substitution, single imputation; Widaman, 2006).

### Conclusion

In this study, we showed that effortful control in late adolescence is a significant predictor of educational attainment by age 23, and its associated effect size was comparable to those of high school CGPA and parental education. This finding indicates the importance of self-regulatory skills for success in postsecondary education and suggests that efforts to improve educational attainment may be enhanced by programs that promote the development of self-regulatory skills. To date, research examining the malleability of effortful control through socialization and through exposure to cognitively and emotionally challenging tasks has shown encouraging results in children and adolescents. Dropout prevention programs could include an effortful control reinforcement component that begins early on and continues throughout the high school years as a way to further support the pursuit and completion of higher education. Substantive and lasting improvement in the level of educational attainment in the population is likely to require a combination of strategies that targets not only individual students but also their environment, including family members, schools, community institutions, and governing bodies at the local and national levels. Programs that support the development of self-regulation may prove to be an important part of these efforts.

### References

- Adams, S. J. (2002). Educational attainment and health: Evidence from a sample of older adults. *Education Economics*, 10, 97–109. doi:10.1080/09645290110110227
- Albrecht, C. M., & Albrecht, D. E. (2011). Social status, adolescent behavior, and educational attainment. *Sociological Spectrum*, 31, 114–137. doi:10.1080/02732173.2011.525698
- Allan, N. P., & Lonigan, C. J. (2011). Examining the dimensionality of effortful control in preschool children and its relation to academic and socioemotional indicators. *Developmental Psychology*, 47, 905–915. doi:10.1037/a0023748
- Andersson, H., & Bergman, L. R. (2011). The role of task persistence in young adolescence for successful educational and occupational attainment in middle adulthood. *Developmental Psychology*, 47, 950–960. doi:10.1037/a0023786
- Arnett, J. J. (2000). Emerging adulthood: A theory of development from the late teens through the twenties. *American Psychologist*, 55, 469–480. doi:10.1037/0003-066X.55.5.469
- Beal, S. J., & Crockett, L. J. (2010). Adolescents' occupational and educational aspirations and expectations: Links to high school activities and adult educational attainment. *Developmental Psychology*, 46, 258–265. doi:10.1037/a0017416
- Belfield, C. R., Levin, H. M., & Brookings, I. (2007). *The price we pay: Economic and social consequences of inadequate education*. Washington, DC: Brookings.
- Belsky, J., & Beaver, K. M. (2011). Cumulative-genetic plasticity, parenting and adolescent self-regulation. *Journal of Child Psychology and Psychiatry*, 52, 619–626. doi:10.1111/j.1469-7610.2010.02327.x
- Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, 78, 647–663. doi:10.1111/j.1467-8624.2007.01019.x
- Bowers, E. P., Gestsdottir, S., Geldhof, G. J., Nikitin, J., von Eye, A., & Lerner, R. M. (2011). Developmental trajectories of intentional self-regulation in adolescence: The role of parenting and implications for positive and problematic outcomes among diverse youth. *Journal of Adolescence*, 34, 1193–1206. doi:10.1016/j.adolescence.2011.07.006

- Chatterji, P. (2006). Illicit drug use and educational attainment. *Health Economics*, 15, 489–511. doi:10.1002/hec.1085
- Checa, P., Rodríguez-Bailón, R., & Rueda, M. R. (2008). Neurocognitive and temperamental systems of self-regulation and early adolescents' social and academic outcomes. *Mind, Brain, and Education*, 2, 177–187. doi:10.1111/j.1751-228X.2008.00052.x
- Checa, P., & Rueda, M. (2011). Behavioral and brain measures of executive attention and school competence in late childhood. *Developmental Neuropsychology*, 36, 1018–1032. doi:10.1080/87565641.2011.591857
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255. doi:10.1207/S15328007SEM0902\_5
- Connell, A. M., Dishion, T. J., & Deater-Deckard, K. (2006). Variable- and person-centered approaches to the analysis of early adolescent substance use: Linking peer, family, and intervention effects with developmental trajectories. *Merrill-Palmer Quarterly*, 52, 421–448. doi:10.1353/mpq.2006.0025
- Connell, A. M., Dishion, T. J., Yasui, M., & Kavanagh, K. (2007). An adaptive approach to family intervention: Linking engagement in family-centered intervention to reductions in adolescent problem behavior. *Journal of Consulting and Clinical Psychology*, 75, 568–579. doi:10.1037/0022-006X.75.4.568
- Connell, A. M., Klostermann, S., & Dishion, T. J. (2012). Family Check-Up effects on adolescent arrest trajectories: Variation by developmental subtype. *Journal of Research on Adolescence*, 22, 367–380. doi:10.1111/j.1532-7795.2011.00765.x
- Coombs, R. H., & Landsverk, J. (1988). Parenting styles and substance use during childhood and adolescence. *Journal of Marriage & the Family*, 50, 473–482. doi:10.2307/352012
- Day, J. C., & Newburger, E. C. (2002). *The big payoff: Educational attainment and synthetic estimates of work-life earnings* (Current Population Reports). Washington, DC: U.S. Census Bureau.
- Diamond, A., Barnett, W. S., Thomas, J., & Munro, S. (2007). Preschool program improves cognitive control. *Science*, 318, 1387–1388. doi:10.1126/science.1151148
- Dishion, T. J., & Kavanagh, K. (2003). *Intervening in adolescent problem behavior: A family-centered approach*. New York, NY: Guilford Press.
- Dishion, T. J., & Stormshak, E. A. (2007). *Intervening in children's lives: An ecological, family-centered approach to mental health care*. Washington, DC: American Psychological Association. doi:10.1037/11485-000
- Doan, S. N., Fuller-Rowell, T. E., & Evans, G. W. (2012). Cumulative risk and adolescent's internalizing and externalizing problems: The mediating roles of maternal responsiveness and self-regulation. *Developmental Psychology*, 48, 1529–1539. doi:10.1037/a0027815
- Dubow, E. F., Boxer, P., & Huesmann, L. R. (2009). Long-term effects of parents' education on children's educational and occupational success: Mediation by family interactions, child aggression, and teenage aspirations. *Merrill-Palmer Quarterly*, 55, 224–249. doi:10.1353/mpq.0.0030
- Eiden, R. D., Edwards, E. P., & Leonard, K. E. (2007). A conceptual model for the development of externalizing behavior problems among kindergarten children of alcoholic families: Role of parenting and children's self-regulation. *Developmental Psychology*, 43, 1187–1201. doi:10.1037/0012-1649.43.5.1187
- Eisenberg, N., Zhou, Q., Spinrad, T. L., Valiente, C., Fabes, R. A., & Liew, J. (2005). Relations among positive parenting, children's effortful control, and externalizing problems: A three-wave longitudinal study. *Child Development*, 76, 1055–1071. doi:10.1111/j.1467-8624.2005.00897.x
- Ellis, L. K. (2002). *Individual differences and adolescent psychosocial development* (Unpublished doctoral dissertation). University of Oregon.
- Ellis, L. K., & Rothbart, M. K. (2001). *Revision of the Early Adolescent Temperament Questionnaire*. Poster presented at the 2001 biennial meeting of the Society for Research in Child Development, Minneapolis, MN.
- Ellis, L. K., & Rothbart, M. K. (2005). *Revision of the Early Adolescent Temperament Questionnaire (EAT-Q)*. Manuscript in preparation.
- Fosco, G. M., Frank, J. L., Stormshak, E. A., & Dishion, T. J. (2013). Opening the "black box": Family Check-Up intervention effects on self-regulation that prevents growth in problem behavior and substance use. *Journal of School Psychology*, 51, 455–468. doi:10.1016/j.jsp.2013.02.001
- Ganzach, Y. (2000). Parents' education, cognitive ability, educational expectations and educational attainment: Interactive effects. *British Journal of Educational Psychology*, 70, 419–441. doi:10.1348/000709900158218
- Goldsmith, H. H., Buss, K. A., & Lemery, K. S. (1997). Toddler and childhood temperament: Expanded content, stronger genetic evidence, new evidence for the importance of environment. *Developmental Psychology*, 33, 891–905. doi:10.1037/0012-1649.33.6.891
- Hampson, S. E., Goldberg, L. R., Vogt, T. M., & Dubanoski, J. P. (2007). Mechanisms by which childhood personality traits influence adult health status: Educational attainment and healthy behaviors. *Health Psychology*, 26, 121–125. doi:10.1037/0278-6133.26.1.121
- Hardy, J. B., Shapiro, S., Mellits, E. D., Skinner, E. A., Astone, N. M., Ensminger, M., . . . Starfield, B. H. (1997). Self-sufficiency at ages 27 to 33 years: Factors present between birth and 18 years that predict educational attainment among children born to inner-city families. *Pediatrics*, 99, 80–87. doi:10.1542/peds.99.1.80
- Herzog, A. R., Franks, M. M., Markus, H. R., & Holmberg, D. (1998). Activities and well-being in older age: Effects of self-concept and educational attainment. *Psychology and Aging*, 13, 179–185. doi:10.1037/0882-7974.13.2.179
- Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55. doi:10.1080/10705519909540118
- Karvonen, J. T., Joukamaa, M., Herva, A., Jokelainen, J., Läksy, K., & Veijola, J. (2007). Somatization symptoms in young adult Finnish population: Associations with sex, education level and mental health. *Nordic Journal of Psychiatry*, 61, 219–224. doi:10.1080/08039480701352611
- King, K. M., Lengua, L. J., & Monahan, K. C. (2013). Individual differences in the development of self-regulation during pre-adolescence: Connections to context and adjustment. *Journal of Abnormal Child Psychology*, 41, 57–69. doi:10.1007/s10802-012-9665-0
- King, K. M., Meehan, B. T., Trim, R. S., & Chassin, L. (2006). Marker or mediator? The effects of adolescent substance use on young adult educational attainment. *Addiction*, 101, 1730–1740. doi:10.1111/j.1360-0443.2006.01507.x
- Kirby, D. (2009). Widening access: Making the transition from mass to universal post-secondary education in Canada. *Journal of Applied Research on Learning*, 2, 1–17.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.
- Kristensen, P., Gravseth, H. M., & Bjerkedal, T. (2009). Educational attainment of Norwegian men: Influence of parental and early individual characteristics. *Journal of Biosocial Science*, 41, 799–814. doi:10.1017/S0021932009990228
- Lengua, L. J. (2006). Growth in temperament and parenting as predictors of adjustment during children's transition to adolescence. *Developmental Psychology*, 42, 819–832. doi:10.1037/0012-1649.42.5.819
- Leung, K., Lau, S., & Lam, W.-L. (1998). Parenting styles and academic achievement: A cross-cultural study. *Merrill-Palmer Quarterly*, 44, 157–172.
- MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, 39, 99–128. doi:10.1207/s15327906mbr3901\_4



- Marjoribanks, K. (2005). Family background, academic achievement, and educational aspirations as predictors of Australian young adults' educational attainment. *Psychological Reports*, 96, 751–754. doi:10.2466/pr0.96.3.751-754
- Marsh, H. W., & O'Mara, A. (2008). Reciprocal effects between academic self-concept, self-esteem, achievement, and attainment over seven adolescent years: Unidimensional and multidimensional perspectives of self-concept. *Personality and Social Psychology Bulletin*, 34, 542–552. doi:10.1177/0146167207312313
- Martin, M. A. (2012). Family structure and the intergenerational transmission of educational advantage. *Social Science Research*, 41, 33–47. doi:10.1016/j.ssresearch.2011.07.005
- McCaul, E. J., Donaldson, G. A., Jr., Coladarc, T., & Davis, W. E. (1992). Consequences of dropping out of school: Findings from high school and beyond. *Journal of Educational Research*, 85, 198–207. doi:10.1080/00220671.1992.9941117
- McLeod, J. D., & Kaiser, K. (2004). Childhood emotional and behavioral problems and educational attainment. *American Sociological Review*, 69, 636–658. doi:10.1177/000312240406900502
- Melby, J. N., Conger, R. D., Fang, S.-A., Wickrama, K. A. S., & Conger, K. J. (2008). Adolescent family experiences and educational attainment during early adulthood. *Developmental Psychology*, 44, 1519–1536. doi:10.1037/a0013352
- Mello, Z. R. (2008). Gender variation in developmental trajectories of educational and occupational expectations and attainment from adolescence to adulthood. *Developmental Psychology*, 44, 1069–1080. doi:10.1037/0012-1649.44.4.1069
- Metzler, C. W., Biglan, A., Ary, D. V., & Li, F. (1998). The stability and validity of early adolescents' reports of parenting constructs. *Journal of Family Psychology*, 12, 600–619. doi:10.1037/0893-3200.12.4.600
- Muraven, M. (2010). Building self-control strength: Practicing self-control leads to improved self-control performance. *Journal of Experimental Social Psychology*, 46, 465–468. doi:10.1016/j.jesp.2009.12.011
- Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- Organisation for Economic Co-operation and Development. (2005). *Rationale for creating a global forum on education*. Retrieved from [http://www.oecd.org/document/56/0,2340,en\\_21571361\\_35013845\\_35123640\\_1\\_1\\_1,100.html](http://www.oecd.org/document/56/0,2340,en_21571361_35013845_35123640_1_1_1,100.html)
- Ou, S.-R. (2005). Pathways of long-term effects of an early intervention program on educational attainment: Findings from the Chicago longitudinal study. *Journal of Applied Developmental Psychology*, 26, 578–611. doi:10.1016/j.appdev.2005.06.008
- Pettit, G. S., Yu, T., Dodge, K. A., & Bates, J. E. (2009). A developmental process analysis of cross-generational continuity in educational attainment. *Merrill-Palmer Quarterly*, 55, 250–284. doi:10.1353/mpq.0.0022
- Piehl, T. F., Véronneau, M.-H., & Dishion, T. J. (2012). Substance use progression from adolescence to early adulthood: Effortful control in the context of friendship influence and early-onset use. *Journal of Abnormal Child Psychology*, 40, 1045–1058. doi:10.1007/s10802-012-9626-7
- Posner, M. I., & Rothbart, M. K. (2007). *Educating the human brain*. Washington, DC: American Psychological Association. doi:10.1037/11519-000
- Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Bub, K., & Pressler, E. (2011). CSRP's impact on low-income preschoolers' preacademic skills: Self-regulation as a mediating mechanism. *Child Development*, 82, 362–378. doi:10.1111/j.1467-8624.2010.01561.x
- Robertson, D. L., & Reynolds, A. J. (2010). Family profiles and educational attainment. *Children and Youth Services Review*, 32, 1077–1085. doi:10.1016/j.childyouth.2009.10.021
- Robins, R. W., John, O. P., Caspi, A., Moffitt, T. E., & Stouthamer-Loeber, M. (1996). Resilient, overcontrolled, and undercontrolled boys: Three replicable personality types. *Journal of Personality and Social Psychology*, 70, 157–171. doi:10.1037/0022-3514.70.1.157
- Ross, C. E., & Mirowsky, J. (2006). Sex differences in the effect of education on depression: Resource multiplication or resource substitution? *Social Science & Medicine*, 63, 1400–1413. doi:10.1016/j.socscimed.2006.03.013
- Rothbart, M. K., & Bates, J. E. (1998). Temperament. In N. Eisenberg (Ed.), *Handbook of child psychology, Vol 3. Social, emotional, and personality development* (5th ed., pp. 105–176). Hoboken, NJ: Wiley.
- Rothbart, M. K., Ellis, L. K., & Posner, M. I. (2011). Temperament and self-regulation. In K. D. Vohs & R. F. Baumeister (Eds.), *Handbook of self-regulation: Research, theory, and applications* (2nd ed., pp. 441–460). New York, NY: Guilford Press.
- Rumberger, R. W. (1987). High school dropouts: A review of issues and evidence. *Review of Educational Research*, 57, 101–121. doi:10.3102/00346543057002101
- Ryan, A. K. (2010). The lasting effects of marijuana use on educational attainment in midlife. *Substance Use & Misuse*, 45, 554–597. doi:10.3109/10826080802490238
- Sahdra, B. K., MacLean, K. A., Ferrer, E., Shaver, P. R., Rosenberg, E. L., Jacobs, T. L., . . . Saron, C. D. (2011). Enhanced response inhibition during intensive meditation training predicts improvements in self-reported adaptive socioemotional functioning. *Emotion*, 11, 299–312. doi:10.1037/a0022764
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research*, 8, 23–74.
- Somech, L. Y., & Elizur, Y. (2012). Promoting self-regulation and cooperation in pre-kindergarten children with conduct problems: A randomized controlled trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51, 412–422. doi:10.1016/j.jaac.2012.01.019
- South, S. J., Baumer, E. P., & Lutz, A. (2003). Interpreting community effects on youth educational attainment. *Youth and Society*, 35, 3–36. doi:10.1177/0044118X03254560
- Steinberg, L., Lamborn, S. D., Dornbusch, S. M., & Darling, N. (1992). Impact of parenting practices on adolescent achievement: Authoritative parenting, school involvement, and encouragement to succeed. *Child Development*, 63, 1266–1281. doi:10.2307/1131532
- Stormshak, E. A., Connell, A., & Dishion, T. J. (2009). An adaptive approach to family-centered intervention in schools: Linking intervention engagement to academic outcomes in middle and high school. *Prevention Science*, 10, 221–235. doi:10.1007/s1121-009-0131-3
- Stormshak, E. A., Fosco, G. M., & Dishion, T. J. (2010). Implementing interventions with families in schools to increase youth school engagement: The Family Check-Up model. *School Mental Health*, 2, 82–92. doi:10.1007/s12310-009-9025-6
- Tang, Y.-Y., Ma, Y., Wang, J., Fan, Y., Feng, S., Lu, Q., . . . Posner, M. I. (2007). Short-term meditation training improves attention and self-regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 17152–17156. doi:10.1073/pnas.0707678104
- Taylor, J. L., Hurd, H. D., Seltzer, M. M., Greenberg, J. S., & Floyd, F. J. (2010). Parenting with mild intellectual deficits: Parental expectations and the educational attainment of their children. *American Journal on Intellectual and Developmental Disabilities*, 115, 340–354. doi:10.1352/1944-7558-115.4.340
- Tobiasz-Adamczyk, B., Bartoszewski, E., Brzyski, P., & Kopacz, M. (2007). Long-term consequences of education, working conditions, and health-related behaviors on mortality patterns in older age: A 17-year observational study in Kraków, Poland. *International Journal of Occupational Medicine and Environmental Health*, 20, 247–256. doi:10.2478/v10001-007-0028-y
- Valiente, C., Lemery-Chalfant, K., & Swanson, J. (2010). Prediction of kindergartners' academic achievement from their effortful control and emotionality: Evidence for direct and moderated relations. *Journal of Educational Psychology*, 102, 550–560. doi:10.1037/a0018992

- Valiente, C., Lemery-Chalfant, K., Swanson, J., & Reiser, M. (2008). Prediction of children's academic competence from their effortful control, relationships, and classroom participation. *Journal of Educational Psychology, 100*, 67–77. doi:10.1037/0022-0663.100.1.67
- Van Ryzin, M. J., & Dishion, T. J. (2012). The impact of a family-centered intervention on the ecology of adolescent antisocial behavior: Modeling developmental sequelae and trajectories during adolescence. *Development and Psychopathology, 24*, 1139–1155. doi:10.1017/S0954579412000582
- Véronneau, M.-H., Dishion, T. J., & Connell, A. M. (2013). *Long-term outcomes of the Family Check-Up model in public secondary schools: A randomized clinical trial linking parent engagement to the progression of substance use from early adolescence to adulthood*. Manuscript submitted for publication.
- Véronneau, M.-H., Vitaro, F., Pedersen, S., & Tremblay, R. E. (2008). Do peers contribute to the likelihood of secondary school graduation among disadvantaged boys? *Journal of Educational Psychology, 100*, 429–442. doi:10.1037/0022-0663.100.2.429
- Widaman, K. F. (2006). Best practices in quantitative methods for developmentalists: III. Missing data: What to do with or without them. *Monographs of the Society for Research in Child Development, 71*, 42–64. doi:10.1111/j.1540-5834.2006.00404.x
- Wolfe, R. N., & Johnson, S. D. (1995). Personality as a predictor of college performance. *Educational and Psychological Measurement, 55*, 177–185. doi:10.1177/0013164495055002002
- Zhang, J.-P., Huang, H.-S., Ye, M., & Zeng, H. (2008). Factors influencing the subjective well being (SWB) in a sample of older adults in an economically depressed area of China. *Archives of Gerontology and Geriatrics, 46*, 335–347. doi:10.1016/j.archger.2007.05.006
- Zhou, Q., Main, A., & Wang, Y. (2010). The relations of temperamental effortful control and anger/frustration to Chinese children's academic achievement and social adjustment: A longitudinal study. *Journal of Educational Psychology, 102*, 180–196. doi:10.1037/a0015908

Received July 13, 2012

Revision received December 18, 2013

Accepted December 26, 2013 ■

### E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at <http://notify.apa.org/> and you will be notified by e-mail when issues of interest to you become available!



# Academic Self-Handicapping and Achievement: A Meta-Analysis

Malte Schwinger  
University of Giessen

Linda Wirthwein  
TU Dortmund

Gunnar Lemmer  
University of Marburg

Ricarda Steinmayr  
TU Dortmund

Self-handicapping represents a frequently used strategy for regulating the threat to self-esteem elicited by the fear of failing in academic achievement settings. Several studies have documented negative associations between self-handicapping and different educational outcomes, *inter alia* academic achievement. However, studies on the relation between self-handicapping and academic achievement have yielded heterogeneous results, indicating the need to conduct meta-analytic investigations and to examine the relevance of several potential moderator variables. This meta-analysis integrates the results of 36 field studies with 49 independent effect sizes ( $N = 25,550$ ). A random effects model revealed a mean effect size between self-handicapping and academic achievement of  $r = -.23$  ( $p < .001$ , range:  $r = -.46$  to  $r = .02$ ). Moreover, moderator analyses showed that the type of self-handicapping scale, the school type (elementary, middle, high school, university), the level of mastery goals in the sample, and the reliability of the self-handicapping scale considerably influenced the mean correlation. Based on our findings, we conclude that educational interventions to enhance academic achievement should additionally focus on preventing self-handicapping.

**Keywords:** self-handicapping, academic achievement, meta-analysis

In the context of academic learning, students sometimes experience threats to their self-esteem. These threats are often elicited by the fear of failing in upcoming achievement situations such as an important exam. A common strategy for regulating this kind of self-esteem threat is *self-handicapping*, which has been defined as constructing impediments to performance to protect or enhance one's perceived competence (Berglas & Jones, 1978). Examples of academic self-handicapping include procrastinating, effort withdrawal, and claiming test anxiety or illness (Urduan & Midgley, 2001). There is substantial agreement in the literature that academic self-handicapping has negative effects on important educational processes and outcomes such as motivation and achievement (e.g., Martin, Marsh, & Debus, 2001a; Urduan, Midgley, & Anderman, 1998; Zuckerman, Kieffer, & Knee, 1998). However, findings from field studies on the relation between self-handicapping and achievement have reflected considerable heterogeneity, ranging from nonsignificant (Rhodewalt & Hill, 1995), to moderately

negative (Boon, 2007; Schwinger & Stiensmeier-Pelster, 2012), to large negative correlations (Midgley & Urduan, 1995, 2001). The variability in results has prohibited researchers from drawing a general conclusion concerning the mean effect of self-handicapping on achievement; and this, in turn, makes it difficult to estimate the implications of self-handicapping for educational practice. Moreover, these inconsistencies in the reported findings suggest that the negative consequences of self-handicapping could be more or less pronounced under different circumstances. Leondari and Gonida (2007), for instance, reported that self-handicapping and achievement were more closely related in elementary compared to high-school students. Another potential moderator might be the self-handicapping scale that is used. Questionnaires measuring habitual self-handicapping focus on different constituent elements of self-handicapping (e.g., using the handicap as an excuse, the *a priori* timing of the strategy). As these elements themselves might be differently related to academic achievement, the choice of a self-handicapping scale might already predispose studies to reach dissimilar conclusions about the relation of self-handicapping to achievement (Urduan & Midgley, 2001).

Several reviews on self-handicapping have been published, each of them addressing a special feature of this wide-spread phenomenon such as the predictive roles of gender (Hirt & McCrea, 2009) or achievement goals (Urduan & Midgley, 2001). Surprisingly, however, there has been no meta-analytic investigation to date of the relation between self-handicapping and achievement in the academic domain. In this article, we present the first meta-analysis on this topic as we seek (a) to provide an empirically sustained estimation of the mean correlation between self-handicapping and

---

This article was published Online First February 10, 2014.

Malte Schwinger, Department of Psychology, University of Giessen, Giessen, Germany; Linda Wirthwein, Department of Psychology, TU Dortmund, Dortmund, Germany; Gunnar Lemmer, Department of Psychology, University of Marburg, Marburg, Germany; Ricarda Steinmayr, Department of Psychology, TU Dortmund, Dortmund, Germany.

Malte Schwinger and Linda Wirthwein contributed equally to this work.

Correspondence concerning this article should be addressed to Malte Schwinger, Department of Psychology, University of Giessen, Otto-Behagel-Str. 10F, 35394 Giessen, Germany. E-mail: malte.schwinger@psychol.uni-giessen.de

achievement in academic settings outside the laboratory and (b) to investigate the impact of several presumed moderators of this relation.

### The Nature of Self-Handicapping

According to Berglas and Jones (1978), self-handicapping is defined as “any action or choice of performance setting that enhances the opportunity to externalize (or excuse) failure and to internalize . . . success” (p. 406). The impetus for self-handicapping is uncertainty about one’s ability, including anticipated threats to one’s self-esteem (Berglas & Jones, 1978; Snyder & Smith, 1982). The postulated protective function of self-handicapping on self-esteem takes advantage of the discounting and augmentation principles of attribution (Kelley, 1971). In the event of failure, the presence of an impediment offers individuals the opportunity to shift attributions for poor performance from low ability (e.g., “I failed the exam because I’m stupid”) to the handicap (e.g., “I failed the exam because I didn’t sleep well last night”). By doing this, ability will be discounted as a causal attribution, and one’s image of competence as well as one’s self-esteem will be buffered. If the individual surprisingly succeeds, attributions to high ability will be augmented because the individual performed well despite the handicap (Tice, 1991).

An important distinction in the literature has been drawn between *behavioral* and *claimed* self-handicapping (Arkin & Baumgardner, 1985; Leary & Shepperd, 1986). Behavioral self-handicapping implies an active acquisition of an impediment, such as drug abuse (Berglas & Jones, 1978), decreased practice time (Baumeister, Hamilton, & Tice, 1985), or choice of debilitating performance settings (Rhodewalt & Davison, 1986). By contrast, claimed self-handicappers only report the presence of obstacles. For example, they claim to suffer from test anxiety (Smith, Snyder, & Handelsman, 1982), physical symptoms (Smith, Snyder, & Perkins, 1983), or a bad mood (Baumgardner, Lake, & Arkin, 1985). These two self-handicapping modes differ from one another in terms of cost-benefit analyses (Hirt, Deppe, & Gordon, 1991). On the one hand, behavioral handicaps are more credible because they are more convincingly tied to performance than claimed ones. For the same reason, however, behavioral handicaps are more costly. On the other hand, claimed handicaps, such as reports of test anxiety, also serve as an excuse for failure but do not necessarily decrease one’s chances of being successful as behavioral handicaps do (Hirt et al., 1991; Leary & Shepperd, 1986; Zuckerman & Tsai, 2005). The conceptually meaningful distinction between claimed and behavioral self-handicapping notwithstanding, the majority of questionnaires designed for assessing habitual self-handicapping in the academic field clearly emphasize behavioral forms of handicapping and do not distinguish between claimed and behavioral self-handicapping.

### The Relation Between Academic Self-Handicapping and Achievement

A theoretical framework for the relation between academic self-handicapping and achievement is provided by the Self-Handicapping and Self-Regulation Cycle (Rhodewalt & Tragakis, 2002; Rhodewalt & Vohs, 2005). In this model, distal motives such as uncertain self-conceptions of competence or low self-

esteem lead to decreased performance expectancies for upcoming tests. The lowered expectancies then serve as proximal motives to use self-handicapping for self-protection. Rhodewalt and Tragakis (2002) assumed that self-handicappers are primarily concerned about their self-worth and less about their actual performance. This skewed focus leads people to choose handicaps that—although effective in protecting their self-esteem—are really detrimental to their performance, such as procrastinating or drinking before an exam. The impaired performance has recursive effects on one’s self-perceptions of ability, thus reinitializing a new cycle of a threatened self-image, self-protection through self-handicapping, and lowered performance<sup>1</sup> (Zuckerman et al., 1998). Although the authors acknowledged that the degree to which performance is impaired may depend on the kind of handicap, they suggested that chronic self-handicapping has long-term negative effects on achievement.

A lot of studies on self-handicapping have been conducted in experimental laboratory settings. Although this kind of research has provided valuable insights into the structure and mode of action of self-handicapping, findings from laboratory studies cannot be generalized to real-world classroom settings in school or college. Thus, laboratory studies do not contribute to the primary purpose of our meta-analysis, which is to unravel the mean association between self-handicapping and achievement in academic settings in order to estimate the severity of the problem for everyday educational practice. Consequently, to be included in the following review, studies (a) must have been conducted in a field setting rather than in a laboratory setting, (b) must have included a sample comprised of school (i.e., elementary, middle, or high school) and/or college or university students, and (c) must have reported a correlation between a self-handicapping questionnaire and a measure of academic achievement (e.g., grade point average [GPA], test scores).

Empirical studies conducted in the academic field with school or university students have reported quite inconsistent results. To categorize the respective effect sizes, we used the reference points for evaluating influences on academic achievement provided by Hattie (2009), who considered  $r = .10$  to be small,  $r = .20$  to be moderate, and  $r = .30$  to be large effects. According to these guidelines, some studies have revealed only a small relation between self-handicapping and achievement of  $r = -.08$  (Urdu, 2004) or  $r = -.07$  (Wesley, 1994). In most cases, however, the two constructs were correlated to a moderately negative degree. In a sample of undergraduate psychology students, Elliot and Church (2003) found self-handicapping to be negatively associated with exam performance ( $r = -.15$ ). Further examples of moderate effect sizes were obtained by Martin, Marsh, and Debus (2001b;  $r = -.19$ ); Zuckerman et al. (1998;  $r = -.20$ ); and McCrea and

<sup>1</sup> The conceptual focus of this article is on the directional effects from self-handicapping to achievement. Given the reciprocal character of the relation between self-handicapping and achievement, we should have included only longitudinal studies in our meta-analysis in which earlier measures of self-handicapping determined later assessments of performance. Due to the small number of longitudinal studies, however, we decided to base the present meta-analysis on both longitudinal and cross-sectional studies. Throughout the article, we thereby seek to maintain the theoretical perspective of directional effects of self-handicapping on achievement, but we are also cautious about our use of causal wording as we acknowledge the correlational design of most studies reviewed here.



Hirt (2001;  $r = -.23$ ). Contrary to these results, other authors have reported fairly larger correlations of, for instance,  $r = -.40$  (Midgley & Urdan, 2001),  $r = -.38$  (Gadbois & Sturgeon, 2011),  $r = -.33$  (Shih, 2005), and  $r = -.38$  (Midgley & Urdan, 1995). In sum, findings have been characterized by substantial heterogeneity with an apparent overlap of moderately negative effect sizes. At first glance, there are several explanations for these heterogeneous correlations.

### Moderator Variables

The studies that have been conducted on the relation between self-handicapping and academic achievement have differed in a number of parameters such as the questionnaires used for assessing self-handicapping, participants' age, the indicators for achievement, and so forth. We now elaborate on how differences in some of these aspects may produce theoretically plausible effects on the magnitude of the correlation between self-handicapping and achievement.

### Self-Handicapping Questionnaire

One of the most apparent differences among studies examining the relation between self-handicapping and achievement is in the questionnaires that have been used to assess the participants' self-handicapping. With some exceptions, researchers have mainly relied on either the Academic Self-Handicapping Scale (ASHS; Midgley & Urdan, 1995; Urdan et al., 1998) or the Self-Handicapping Scale (SHS; Jones & Rhodewalt, 1982). Whereas the six-item ASHS has been used in essentially the same form over time (Urdan & Midgley, 2001), the SHS has been used both in its original form with 25 items and in short versions with 10 items (Strube, 1986) and 14 items (Rhodewalt, 1990; Zuckerman et al., 1998), respectively. Despite some overlap, the ASHS and the SHS show considerable differences in their operationalization of self-handicapping. In our view, the ASHS items were developed in a straightforward manner in conjunction with existing theory. Urdan and Midgley (2001) denoted three features necessary for a valid self-handicapping item. It has to include the handicapping behavior (e.g., effort withdrawal), the reason for this behavior (e.g., to use low effort as an excuse), and the a priori timing of the strategy (e.g., low effort as an excuse is installed *before* failure occurs). All ASHS items have been formulated in line with these recommendations (e.g., "Some students put off doing their school work until the last minute so that if they don't do well on their work, they can say that is the reason. How true is this of you?"). By contrast, items from Jones and Rhodewalt's (1982) SHS do not fully represent these criteria. Many items just ask for a behavior that has the potential to be a handicapping behavior, thereby leaving the reason for it open (e.g., "I tend to put things off until the last moment"; "I am easily distracted by noises or my own daydreaming when I try to read"). Likewise, another set of SHS items emphasize a person's tendency to search for excuses in the case of failure. However, the a priori timing of installing the excuse before failure occurs is not integrated into these items (e.g., "I tend to make excuses when I do something wrong"). Altogether, the SHS items are only partially in line with Urdan and Midgley's (2001) required features of a valid self-handicapping item, and the criteria that they fulfill are not consistent across all SHS items.

In light of the differences described above, we believe that choosing either the ASHS or the SHS can be responsible for dissimilar correlations between self-handicapping and achievement. So which kind of relation would be expected for which kind of questionnaire? We assume that agreeing with statements from the ASHS will have more deleterious effects on students' performance than agreeing with items from the SHS. This seems reasonable to assume because the SHS does not necessarily assess individuals' tendencies to self-handicap but rather measures some kind of undifferentiated avoidance behavior. High scores on the SHS might thus be justified by a number of reasons that may not have been considered when the items were formulated. Putting things off until the last moment, for instance, might reflect a stress-induced kind of time management that does not necessarily lead to lower achievement (Steel, 2007). Likewise, tending to use excuses *after* failure may be just an adaptive self-protective reaction that says nothing about one's tendency to establish excuses *before* an important test or exam.

Taken together, we think that the maladaptive effects of self-handicapping on achievement accumulate with each step of the self-handicapping process. Whereas executing a behavior that has the potential to handicap one's test score may already be detrimental, it is even worse when it is used as an excuse as well as when it is implemented right before the respective test situation. A person's behavior must agree with all three criteria to paint the picture of a self-handicapping person who feels threatened by anticipated failure; who invests more time in ruminating about him- or herself than in learning; and who is prone to entering a vicious cycle of low performance, increased self-esteem threat, and repeated self-handicapping again. However, this form of academic self-handicapping is exclusively represented by the ASHS, so it would be reasonable to assume that higher negative correlations between self-handicapping and achievement would be found in studies using this measure compared to the SHS.

In addition to the ASHS and SHS, the Self-Sabotage subscale of the Motivation and Engagement Scale (MES; see Liem & Martin, 2012, for an overview) has also been used in some studies. The MES was designed to represent 11 cognitive and behavioral dimensions relevant to motivation and engagement as proposed by Martin (2007) in his Motivation and Engagement Wheel. In the MES, self-sabotage (in fact, often termed "self-handicapping" in more recent work) is measured with four items that were adapted from either the ASHS or the short version of the SHS (Strube, 1986). In fact, however, all items fulfill the requirements for a valid self-handicapping item outlined by Urdan and Midgley (2001); thus, we assumed that the ASHS and the self-sabotage scale would yield very similar correlations with academic achievement.

### School Type and Age

In order to provide teachers and practitioners with information about the possible starting points of such vicious cycles, it is crucial to investigate potential age-related differences in the effects of self-handicapping on achievement. However, this topic has seldom been addressed in the literature. To our knowledge, Leonardari and Gonida's (2007) study is the only one that has provided a direct comparison between students of different ages. They found a moderately negative correlation between self-handicapping and



mathematics achievement in elementary and junior high school students, whereas there was a null relation in senior high school students. Descriptive inspections of other studies have supported these results. At first glance, studies examining samples of college or university students seem to have reported lower correlations (e.g., Martin et al., 2001b) compared to studies investigating samples of younger school students (e.g., Midgley & Urdan, 2001). This overview of empirical evidence suggests that a meta-analytic investigation could also reveal meaningful age-related differences.

These empirical findings notwithstanding, theoretical justifications for age differences have been scarce in the self-handicapping literature. Zuckerman et al. (1998) even hypothesized that chronic self-handicappers should show an accelerated decline in academic achievement as the vicious circle they engage in results in changes in academic achievement for the worse. However, the results depicted above contradict this hypothesis. One possible explanation for the findings reported above can be found in the different grading structures that are used in elementary and middle school. In the grading systems used by teachers of younger students, more emphasis is placed on soft factors such as motivation to learn, classroom behavior, and so forth than in the systems used by teachers of older students (McMillan, 2001; Remesal, 2011). Thus, effort withdrawal or other kinds of handicapping behaviors could contribute a particularly large amount of negative weight toward the grades of younger students. This could lead to an inflated correlation between self-handicapping and achievement in younger students.

Another possible explanation might come from developmental differences in students' self-evaluations of their ability. Young children do not have a clearly differentiated definition of academic competence (Marsh, 1992; Stipek & Mac Iver, 1989); thus, they might interpret failure in a specific domain to mean that they are generally less capable in school. Such generalization processes could induce a self-reinforcing cycle of lowered success expectancies and self-handicapping in all domains, which, in turn, might accumulate into more deleterious effects on performance compared to students with more differentiated self-perceptions of their ability. Overall, we expected to identify age-related differences in the correlation between self-handicapping and achievement in the present meta-analysis. Those differences might be explained by developmental issues as described above.

## Gender

In numerous studies, men have been found to show higher scores on self-handicapping questionnaires than women (e.g., Midgley & Urdan, 1995, 2001; Urdan et al., 1998). These results and also the fact that women use behavioral forms of self-handicapping less frequently than men represent very robust findings in self-handicapping research (see Hirt & McCrea, 2009, for a review). Whereas all people prefer claimed over behavioral self-handicapping, this tendency is quite a bit more observable in women than in men (Hirt et al., 1991). Researchers have struggled for a long time to explain these mean differences in behavioral self-handicapping. To date, the most prominent assumption stresses the differential valuing of effort. McCrea, Hirt, Hendrix, Milner, and Steele (2008) introduced the Worker scale, an instrument that assesses the extent to which an individual sees him/

herself as a hard worker and personally values these characteristics. The authors found that women tend to score higher on this measure. Moreover, the Worker scale partially mediated the relation between gender and behavioral self-handicapping (i.e., women showed higher scores on the Worker scale but lower scores on behavioral self-handicapping).

Only a few studies have reported correlations between self-handicapping and achievement separately for women and men. Whereas McCrea et al. (2008) reported similar correlations between self-handicapping and GPA for women and men, Wesley (1994) found a substantially higher correlation for men ( $r = -.36$ ) compared to women ( $r = -.23$ ). Due to this small database, we deemed it important to meta-analytically examine the potential moderating effect of gender. From a conceptual standpoint, the Worker scale findings by McCrea et al. (2008) may be helpful for illuminating the extent to which gender moderates the correlation between self-handicapping and achievement. Women seem to be smarter when choosing a self-protection strategy. With respect to self-handicapping, they tend to rely on claimed rather than behavioral self-handicapping (Hirt & McCrea, 2009). However, because all questionnaires included in this meta-analysis put more emphasis on behavioral forms of self-handicapping, this aspect alone was not expected to produce a moderating effect of gender. It was thus more important to consider whether women would self-handicap in a smarter way than men even when they handicapped behaviorally, too. For example, women may be more sensitive to having an adequate degree of effort withdrawal, thereby not reducing their effort more than necessary. If true, this would be reflected in a substantially lower correlation between self-handicapping and academic achievement for women compared to men.

## Achievement Goals

Achievement goals refer to the reasons why people engage in achievement-related situations. Recent frameworks mainly differentiate four different achievement goals (Elliot & McGregor, 2001): mastery-approach goals (enhancing task-based or intrapersonal competence), mastery-avoidance goals (avoiding task-based or intrapersonal incompetence), performance-approach goals (demonstrating high ability or competence to others), and performance-avoidance goals (avoiding appearing incompetent to others). Several studies have found positive links between self-handicapping and performance-avoidance goals (e.g., Elliot & Church, 2003; Leondari & Gonida, 2007) but negative links between self-handicapping and mastery goals (e.g., Martin et al., 2001b; Midgley & Urdan, 2001).

However, students can pursue multiple goals simultaneously. Schwinger and Stiensmeier-Pelster (2011) reported that the additional endorsement of mastery-approach goals buffered the relation between performance-avoidance goals and self-handicapping most likely because mastery goals decreased the self-esteem-threatening effect of anticipated failure by suggesting attributions of failure that emphasize controllable factors such as low effort. The relation between self-handicapping and academic achievement may be influenced in a similar way: The more students also pursue mastery goals, the more they might attribute ongoing failure to controllable factors that might help them to find a way out of the self-handicapping cycle. Based on this reasoning, we as-



sumed that correlations between self-handicapping and achievement would be lower for highly mastery-oriented students.

### **Self-Esteem, Self-Efficacy, and Academic Self-Concept**

Cognitive and affective self-evaluations also serve as determinants of self-handicapping, whereby positive views of the self are generally associated with less of a need to handicap (Schwinger & Stiensmeier-Pelster, 2012; Zuckerman et al., 1998). Parallel to the reasoning for achievement goals, we assume that positive self-views suggest more adaptive attributions of one's performance, and thus, students with positive self-views should subsequently self-handicap for a shorter time period and/or less strongly. Consequently, we proposed that the negative effects of self-handicapping on achievement would be lower for students with rather positive compared to negative self-perceptions.

### **Level of Achievement**

Most handicapping behaviors are supposed to interfere with deep and successful learning. It might make a difference, however, if the handicapping student has usually performed well or poorly in the past. As previous knowledge and performance in a given domain are among the most powerful predictors of further achievement (Hattie, 2009; Steinmayr & Spinath, 2009), it seems reasonable to assume that previously low-achieving students will further fall behind after using self-handicapping strategies. By contrast, usually high-achieving but occasionally handicapping students will be able to draw on their accumulated knowledge, which may result in only a marginal drop—if there is any drop at all—in school performance. Thus, we deemed it possible that the relation between self-handicapping and achievement would be higher for low-achieving compared to high-achieving students.

### **Origin of the Sample**

As self-handicapping represents a strategy for regulating one's self-esteem, it is important to consider how individual self-esteem is psychologically construed. Self-construal is supposed to differ as a function of individualism-collectivism. In individualistic cultures, the self is construed in independent terms as a separate, distinct entity, and the main task of the person is to "stand out" by distinguishing oneself from others through self-sufficiency and personal accomplishment. In collectivistic cultures, the self is construed in interdependent terms as a connected relational entity, and the main task of the person is to "fit in" by maintaining interpersonal relationships and group harmony (Markus & Kitayama, 1991). In individualistic cultures such as the United States, the attainment of positive outcomes is emphasized and valued, whereas in collectivistic cultures such as South Korea and Russia, avoiding negative outcomes is emphasized and valued (Elliot, Chirkov, Kim, & Sheldon, 2001). Drawing on these findings, we assume that self-handicapping is seen more positively in collectivistic cultures, probably yielding to a more benign evaluation of self-handicapping persons. More benign performance evaluations, in turn, might reduce the negative relation between self-handicapping and academic achievement.

### **Ethnicity**

Several studies have addressed the question of whether students belonging to cultural minorities (e.g., African American) might be more prone to self-handicapping. Urdan and Midgley (2001) argued that stereotype threat among minorities makes self-handicapping more likely in these cultural groups. For example, when African American students are concerned with appearing academically able, the threat of fulfilling a negative stereotype about African Americans' low academic ability is activated and there is a greater need to avoid appearing academically unable than there is for European American students, who have no such stereotype. Such processes may result in higher handicapping for students from any stereotype-threatened cultural group (Urdan et al., 1998). Even more important, however, stereotype attitudes may also be activated in the teachers. There is ample evidence that such stereotype attitudes bias teachers' interpretations of both the adaptive and maladaptive behaviors of their students (Gunderson, Ramirez, Levine, & Beilock, 2012). In this regard, teachers might see their stereotype fulfilled in self-handicapping minority students and may therefore assign immoderately bad grades to these students. To our knowledge, only a few studies (e.g., Midgley, Arunkumar, & Urdan, 1996) have reported correlations between self-handicapping and achievement separately for different ethnic groups, so we deemed it necessary to investigate this issue meta-analytically here.

### **Achievement Indicators**

The relation between self-handicapping and achievement may depend on the achievement indicators considered. Meta-analyses on related constructs have sometimes yielded substantial differences for distinct measures of achievement. In his meta-analysis on procrastination, Steel (2007) reported a significantly weaker negative relation between procrastination and overall GPA compared to course-specific GPA. Investigating the meta-analytic effects of achievement goals, Wirthwein, Sparfeldt, Pinquart, Wegerer, and Steinmayr (2013) found significantly higher negative correlations between performance avoidance goals and standardized test scores as well as specific exam grades compared to GPA (see also Huang, 2012). For mastery-approach and performance-approach goals, lower correlations emerged for standardized achievement test scores compared to other achievement indicators. Similar results were found in several meta-analyses on the academic self-concept (Hansford & Hattie, 1982; Huang, 2011). For our meta-analysis, we decided that it would be crucial to distinguish between standardized achievement tests and teacher-assigned grades. With respect to the latter, it would be further important to differentiate between a specific grade on one exam and/or in one school subject from averaged measures such as GPA. As stated above, grades are biased to a certain degree by teachers' perceptions. With respect to self-handicapping, we suggest that teachers do not appreciate finding handicapping behavior in their classes. Specifically, they might interpret such behavior as laziness, and this might lead them to lower the self-handicapper's grade (Covington & Omelich, 1979). Therefore, we predicted that the negative relation between self-handicapping and achievement would be higher when grades were used as achievement indicators compared to standardized test scores.

## Publication Status

Significant findings are more likely to be accepted for publication than nonsignificant ones (Ferguson & Brannick, 2012). This publication bias may be even larger for constructs for which rather high correlations with respective outcomes are theoretically presumed. Given that self-handicapping is already defined as a self-impeding and performance-inhibiting strategy, it can be extremely difficult for researchers to publish data showing null relations between self-handicapping and achievement. We therefore sought to establish whether the respective correlations would be significantly different for published versus unpublished studies.

## Concurrent Versus Prospective Measurement

Several authors have noted that the relation between self-handicapping and achievement is probably reciprocal (Covington, 1992; McCrea et al., 2008; Zuckerman et al., 1998). That is, fear of failing on a test leads to self-handicapping, which, in turn, decreases one's performance. This low performance even further enhances self-doubts about mastering the next test, and these self-doubts lead to the necessity to self-handicap again. Conceptually, we are interested in the causal effects of self-handicapping on achievement and not vice versa. Due to the small number of longitudinal studies, however, we had to base the present meta-analysis on both longitudinal and cross-sectional studies. However, if we were to find that the prospective effects of self-handicapping on achievement were substantially lower compared to concurrent correlational effects, this may provide some hints that at least some amount of the correlation can be attributed to the directional effect of achievement on self-handicapping.

## Reliability of the Self-Handicapping Scale

Another methodological aspect frequently considered in meta-analyses is the quality of the individual studies. However, because of its subjectivity, considering a study's quality is seen as quite controversial in the literature (e.g., Jüni, Witschi, Bloch, & Egger, 1999). Hence, we focused on the reliability (Cronbach's alpha) of the self-handicapping scale as a selected objective criterion for the methodological quality of a study. In this context, we assumed that the higher the reliability, the higher the correlation between self-handicapping and achievement would be.

## Specificity of the Self-Handicapping Measurement

Several constructs related to academic self-handicapping have been found to be more predictive of achievement when they were measured in a domain-specific manner. Hansford and Hattie (1982), for instance, revealed mean correlations between domain-specific self-concepts and achievement that were almost twice as high as those between general measures of self-concept and achievement. Similar results were obtained by Huang (2011) and Valentine, DuBois, and Cooper (2004). With respect to achievement goals, Huang (2012) reported that domain-specific measures of performance-avoidance goals were more strongly related to achievement than general ones. Based on these findings, we assumed that the relation between self-handicapping and achievement would be higher when self-handicapping was measured domain-specifically.

## Achievement Domain

As an extension of the issue of domain-specificity, we were also interested in whether self-handicapping would be more likely to produce deleterious effects on achievement in certain school domains. In particular, we sought to examine differences between mathematics and language subjects, as we presumed a distinct self-handicapping potential for math. In mathematics, students might experience tasks as solvable or not, which would serve as indirect feedback concerning intelligence for the student. Therefore, the degree to which a student would attribute failure to internally stable reasons (e.g., low intelligence) in mathematics may be higher compared to other subjects (e.g., English). Ability attributions of failure can lead to self-esteem threat, which, in turn, might enhance the probability that one will use self-handicapping strategies. Haag and Götz (2012) reported that students perceived the characteristics of mathematics completely differently than those of English. Math was especially characterized as having a high potential to induce self-threatening events and, thus, a high potential for self-handicapping. Most notably, students rated it as a subject in which one needs to be intelligent because diligence alone is not enough to obtain good grades. Unlike English, moreover, the right solutions for tasks in math were rated as clear and without ambiguity. Finally, math was characterized as more effortful and more difficult than English. Because self-handicapping thus seems to be used more frequently in math, we suggest that the risk of becoming a chronic self-handicapper in math should also be considerably higher. Habitual self-handicapping in a domain, however, increasingly undermines achievement, and this is why we predicted that there would be larger negative correlations between self-handicapping and achievement in math compared to language subjects. Supporting these ideas, Huang's (2012) meta-analysis revealed that performance-avoidance goals displayed the highest negative effects on achievement in math, whereas Wirthwein et al. (2013) found that the achievement indicator was just a relevant moderator for performance-approach goals.

## Domain Matching

The relevance of assessing all variables on the same level of specificity has already been discussed elsewhere (e.g., Baranik, Barron, & Finney, 2010). In line with these previous studies, we expected the correlations between self-handicapping and achievement to be higher when the levels of specificity were matched, that is, when both variables were assessed on a global or domain-specific level (e.g., global self-handicapping and GPA or self-handicapping in math and achievement in math) compared to a "mismatch" (e.g., global self-handicapping and achievement in math). In the related field of achievement-goal research, Wirthwein et al. (2013) and Huang (2012) found some evidence that this "domain matching" moderated the associations between mastery goals and achievement as well as between performance-avoidance goals and achievement.

## The Present Research

There are several reasons that support why a meta-analysis on the relation between self-handicapping and achievement is clearly needed. First, the heterogeneity in correlations between self-



handicapping and achievement has not yet been addressed meta-analytically. Likewise, there has been no literature review that has addressed this topic in more detail. Existing reviews have rather been dedicated to determinants of self-handicapping such as gender (Hirt & McCrea, 2009), claimed versus behavioral handicapping (Leary & Shepperd, 1986), or achievement goals (Urdañ & Midgley, 2001). As a consequence, little is known about (a) the mean effect size of the relation between self-handicapping and achievement and (b) which factors are the most relevant moderators of this effect. A further justification for the current meta-analysis is that it will help to rank self-handicapping as a determinant of school performance. In his synthesis of meta-analyses, Hattie (2009) ranked the most common predictors of school performance by their mean effect sizes, thereby providing guidance on the importance of each construct for educational practice. Certainly, more research efforts will be put toward psychological intervention programs for the variables with higher effect sizes. Unraveling the mean effect of self-handicapping on achievement will allow us to compare this effect with those of other important school performance predictors and to estimate the relative necessity of designing intervention programs against self-handicapping.

In this article, we present the first meta-analysis on the relation between self-handicapping and achievement in the academic domain. We sought to explore the mean effect size of the correlation between self-handicapping and achievement as well as to explain the heterogeneity in empirical findings by identifying significant moderators. Specifically, we examined the moderating impact of (a) different self-handicapping questionnaires; (b) school type; (c) different sample characteristics; (d) achievement goals; (e) self-esteem, self-efficacy, and academic self-concept; (f) level of achievement; (g) methodological aspects (publication type, time of measurement, reliability of the self-handicapping scale); (h) different achievement indicators; (i) specificity of self-handicapping measurement; (j) different achievement domains; and (k) domain matching.

## Method

### Literature Search and Coding

We systematically searched electronic databases (i.e., PsycINFO, ERIC, Google Scholar, Web of Knowledge, ProQuest Dissertations & Theses, OpenGrey, NTIS, PSYINDEX) for abstracts that contained one of the search terms self-handicapping, self-sabotage, self-deception, self-defeating behavior, safeguarding, self-deceiving, self-impairment, effort withdrawal, self-impediment, or self-hindering and either the term achievement or the term performance. We included studies up to August 2013. This search led to a maximum of 366 abstracts to be checked. We also used cross-referencing to identify relevant studies. Studies were then included in the meta-analysis if (a) correlations between self-handicapping and academic achievement were specified (excluding achievement in sports), (b) the research was conducted with self-report measures, (c) the sample comprised school or university students, and (d) the study was written in English or German. Moreover, we personally contacted known authors in the field and asked them for unpublished studies or unpublished data that matched the aforementioned criteria. By applying these criteria,

we identified 36 studies and 49 effect sizes (see Table 1 for a list of the included studies with selected descriptive statistics).

Table 2 shows the moderator variables that we considered, the coding of the categories, and the respective kappa coefficients. We computed Cohen's kappa to calculate the coding reliability of the variables (Cohen, 1992). Therefore, a second coder additionally categorized 10 of the included studies (i.e., about 34%). Kappas between .61 and .80 were classified as substantial and between .81 and 1.00 as excellent (Landis & Koch, 1977). As can be seen, the coding reliabilities of each item were at least substantial with kappas ranging from .68 to 1.00. With two exceptions regarding the categories school type ( $\kappa = .68$ ) and questionnaire ( $\kappa = .74$ ), all kappas could be classified as excellent.

### Effect Size Calculation and Analyses of Effect Sizes

We considered the Pearson product moment correlation coefficient as the effect size statistic for our meta-analysis. Outliers were defined as correlations that were more than 2 *SDs* above or below the mean of the correlation (this was true for just two correlations). According to Lipsey and Wilson (2001), we winsorized these two outliers to a less extreme value (2 *SDs*). If multiple effect sizes were reported in one study (e.g., different academic achievement indicators), we combined them into a single effect size using Fisher's *Z* scores to avoid dependency in the data (Lipsey & Wilson, 2001). If results in one study were reported for different groups (e.g., gender, age), they were handled as two distinct and independent results.

With reference to the total effect, we chose a priori to integrate effect sizes by using the random effects model (REM; Hedges & Vevea, 1998), as its theoretical postulate allows the individual true effects to differ. We used a restricted maximum likelihood (REML) estimator of the variance of the true effects. As demonstrated by Monte Carlo simulations (Viechtbauer, 2005), this estimator is efficient and has few biases. Each correlation coefficient was transformed into a Fisher's *Z* score, and each effect size was weighted according to the REM by the inverse of the sum of the sampling variance and the estimated variance between the true effects (Lipsey & Wilson, 2001). To compute an estimate for the mean of the true effects of the individual studies, each effect size was multiplied by its weight, and the sum of these products was divided by the sum of the weights. To gain further insights into the heterogeneity of the effects, we used Cochran's *Q*-Test for homogeneity (see Hedges & Olkin, 1985) and the  $I^2$  statistic (Higgins & Thompson, 2002). Finally, all weighted mean effect sizes and corresponding confidence intervals were converted back to Pearson product moment correlation coefficients.

Additionally, we used procedures to assess whether the results could have been affected by publication bias (Rothstein, Sutton, & Borenstein, 2005), which refers to an overestimation of the average true effect due to the circumstance that published studies have larger effects than unpublished documents. First, we inspected funnel plots (Light, Singer, & Willett, 1994), which plot the individual effect sizes against their corresponding standard errors. An asymmetric distribution of the effect sizes around the estimated mean of the true effect can signal that the sample of the included studies is potentially biased. Second, the funnel plots were statistically tested for asymmetry with a rank correlation test (Begg &

Table 1  
Included Studies With Selected Descriptive Statistics

Study	N	School status	Gender	Questionnaire	AI
Boon (2007)	879	3		1	2
Clarke et al. (2013)	85	4	1	1, 2, 5, 6	3
Cocorada (2011)	232	4	1	6	1
De Castella et al. (2013)	643	3	2	1	4
Elliot & Church (2003)	181	4	2	2	1, 2
Feick & Rhodewalt (1997)	121	4	1	2	3
Gadbois (2013)	56	4		1	1, 4
Gadbois (2013)	43	4		1	1, 4
Gadbois & Sturgeon (2011)	209	4	2	1	3
Kleitman & Gibson (2011)	177	1	1	1	2
Leondari & Gonida (2007), Study 1	255	1		1	4
Leondari & Gonida (2007), Study 2	249	2		1	4
Leondari & Gonida (2007), Study 3	198	3		1	4
Martin (2003), Study 1	269	3	1	5	1, 2
Martin (2003), Study 2	1,600	3	1	5	2
Martin & Hau (2010), Study 1	528		1	5	1
Martin & Hau (2010), Study 2	6,366		1	5	2
Martin et al. (2001a)	328	4	2	4	1
Martin et al. (2001b)	584	4	2	4	1
Martin et al. (2003)	291	4	2	4	4
Martin et al. (2013)	969	3	1	5	2
McCrae (2013a)	552	4	1	2	1, 2
McCrae (2013a)	125	4	2	2	1, 2, 3
McCrae (2013b)	57	3	1	2	3, 4
McCrea & Hirt (2001)	158	4	1	6	3
McCrea et al. (2008), Study 1 (Male)	129	4	1	2	1, 2
McCrea et al. (2008), Study 1 (Female)	387	4	2	2	1, 2
McCrea et al. (2008), Study 2 (Male)	454	4	1	2	1, 2
McCrea et al. (2008), Study 2 (Female)	1,035	4	2	2	1, 2
Midgley et al. (1996)	112	2	1	1	1
Midgley & Urdan (1995)	256	2	1	1	1
Midgley & Urdan (2001)	484	2	1	1	4
Murray & Warden (1992)	208	4	2	6	3
Plenty & Heubeck (2011)	1,014	3	1	4	4
Rhodewalt & Hill (1995)	86	4	2	2	3
Schwinger (2013), Study 1	105	3	1	1	4
Schwinger (2013), Study 2	749	3	1	1	4
Schwinger & Kreppold (2012)	1,023	1	1	1	1
Schwinger & Stiensmeier-Pelster (2010)	389	3	2	1	3
Schwinger & Stiensmeier-Pelster (2012), Study 1	613	4	2	1	3
Schwinger & Stiensmeier-Pelster (2012), Study 2	143	4	2	1	3
Shih (2005)	242	1	1	1	1
Thomas & Gadbois (2007)	161	4	2	1	1
Turner et al. (2002)	1,092	1	1	1	4
Urdan (2004)	675	3	1	1	4
Urdan et al. (1998)	528	1	1	1	1
Wesley (1994), Study 1	54	4	1	3	1, 2
Wesley (1994), Study 2	194	4	2	3	1, 2
Zuckerman et al. (1998)	262	4	2	2	1

*Note.* School type: 1 = elementary school, 2 = middle school, 3 = high school, 4 = university; Gender: 1 = ≤59% female, 2 = >59% female; Questionnaire: 1 = Academic Self-Handicapping Scale (Midgley & Urdan, 1995), 2 = Self-Handicapping Scale (Jones & Rhodewalt, 1982), 3 = Short Self-Handicapping Scale (Strube, 1986), 4 = Mixed (Midgley & Urdan, 1995; Strube, 1986), 5 = Motivation and Engagement Scale Self-Sabotage subscale (Liem & Martin, 2012), 6 = Others; AI = academic achievement indicator: 1 = grade point average, 2 = achievement test, 3 = exam grade, 4 = school report-card grade.

Mazumdar, 1994) and a regression test (Egger, Davey Smith, Schneider, & Minder, 1997).

To examine potential moderators, we referred to the meta-analytic mixed effects model (MEM; Hedges & Olkin, 1985; Raudenbush, 2009), which transfers the REM to the fixed values of potential moderators. We analyzed the variability in the effect

sizes due to differences between the categories of the respective potential moderator (e.g., different achievement indicators) with a weighted meta-analytic analogue to the analysis of variance. A statistically significant  $Q_B$ -score implies that the mean effect sizes of the groups or categories of the respective moderator differ by more than sampling error (Lipsey & Wilson, 2001). We referred to



Table 2  
Coding Scheme and Interrater Reliability

Variable	Coding	K
Questionnaire	1 = Academic SHS (Midgley & Urdan, 1995); 2 = SHS (Jones & Rhodewalt, 1982); 3 = Mixture (Midgley & Urdan, 1995; Strube, 1986); 4 = Others	.74
School type	1 = elementary school; 2 = middle school; 3 = high school; 4 = university; 5 = mixed	.68
Gender	1 = ≤59% female; 2 = >59% female	1.00
Achievement goals; self-esteem; self-efficacy; academic self-concept; achievement level in the sample	1 = low; 2 = medium sized/average; 3 = high	1.00
Origin of the sample	1 = United States; 2 = Europe; 3 = Asia; 4 = Australia; 5 = South America; 6 = Canada	.90
Achievement indicator	1 = GPA; 2 = achievement test score (SAT, etc.); 3 = exam grade; 4 = semester/school report-card grade; 5 = mixed	.87
Time	1 = concurrent assessment; 2 = prospective assessment	.87
Specificity self-handicapping	1 = general/school/university; 2 = math and sciences; 3 = others	1.00
Achievement domain	1 = general/school/university; 2 = math and sciences; 3 = languages and other subjects; 4 = mixed	.92
Domain matching	1 = mismatch; 2 = match	1.00

Note. SHS = Self-Handicapping Scale; GPA = grade point average; SAT = Scholastic Aptitude Test.

the principle that nonoverlapping 95% confidence intervals (CIs) indicate a meaningful difference between two effect sizes (see Lipsey & Wilson, 2001) to obtain information on the discrepancy between two specific categories. For continuous variables such as ethnicity or the reliability of the self-handicapping scales, weighted least squares (WLS) meta-regression analyses (see Steel & Kammeyer-Mueller, 2002; Viechtbauer, 2008) were used. One meta-analysis of variance (ANOVA) or meta-regression analysis was conducted for each moderator. In addition, a multiple WLS meta-regression was conducted to identify the relevance of one moderator compared to others.

To obtain all relevant information for the moderator analyses regarding the categories “achievement indicator,” “specificity of the self-handicapping-scale,” “achievement domain,” and “domain matching,” all correlation coefficients in each study were used (and not the averaged correlations). The above-mentioned analyses were conducted with the SPSS macros developed by Lipsey and Wilson (2001) and with the metafor package (Viechtbauer, 2010) for R (R Development Core Team, 2010).

## Results

### Mean Effect Size

The 36 included studies comprised  $k = 49$  independent samples with a total of  $N = 25,550$  participants (range:  $N = 43$  to  $N = 6,366$ ). Publications came from the United States (38.8%), Europe (30.6%), or Australia (20.4%). The studies were published between the years 1992 and 2013. University students comprised 51% of the participants, and 49% were school students.

The random-effects model revealed a mean correlation between self-handicapping and achievement of  $r = -.23$  ( $p < .001$ ; range:  $r = -.46$  to  $r = .02$ ; 95% CI  $[-.25, -.20]$ ;  $k = 49$ ). A forest plot of the included studies can be found in Figure 1. According to the standard already provided, the correlation is of a medium size (Hattie, 2009). Next, we tested whether this mean effect size could

be influenced by publication bias. An inspection of the funnel plot in Figure 2 showed that the individual effects were not asymmetrically distributed around this estimate of the mean of the true effects. This impression was confirmed by statistical tests of funnel plot asymmetry. Neither the rank correlation test (Kendall's  $\tau = 0.01$ ,  $p = .90$ ) nor the regression test ( $z = -0.36$ ,  $p = .72$ ) indicated a funnel plot asymmetry, so there were no indications that the findings were biased.

Cochran's  $Q$ -Test suggested heterogeneity ( $Q = 156.50$ ,  $df = 48$ ,  $p < .001$ ), which means that the individual observed effects differed more than would be expected if sampling error were the only source of variability. Hence, there were differences among the true effects. The amount of total variability between the observed effect sizes that was due to heterogeneity was estimated to be  $I^2 = 70.32\%$ , 95% CI  $[55.72\%, 82.96\%]$ , and could be classified as “high” (Higgins & Thompson, 2002). Subsequently, we tested whether the heterogeneity could be (at least) partially explained by the variables that we considered as potential moderators.

### Moderator Analyses

Table 3 provides an overview of the moderator analyses.<sup>2</sup> Four significant moderator variables were identified. The largest moderator effect was found for “school type” ( $Q_B = 18.83$ ,  $p < .01$ ). The confidence intervals of the categories “elementary school” ( $r = -.29$ ; 95% CI:  $-.35 \leq r \leq -.23$ ) and “middle school” ( $r = -.34$ ; 95% CI:  $-.41 \leq r \leq -.25$ ) compared to “high school” ( $r = -.23$ ; 95% CI:  $-.26 \leq r \leq -.18$ ) and “university students” ( $r = -.18$ ; 95% CI:  $-.21 \leq r \leq -.14$ ) did not overlap, indicating a statistically significant difference between these correlations. Furthermore,

<sup>2</sup> We also examined possible moderating effects of language (German, English), research group (groups of authors/researchers), and year of publication. However, because we found no significant effects, we decided not to report these moderators in the article.

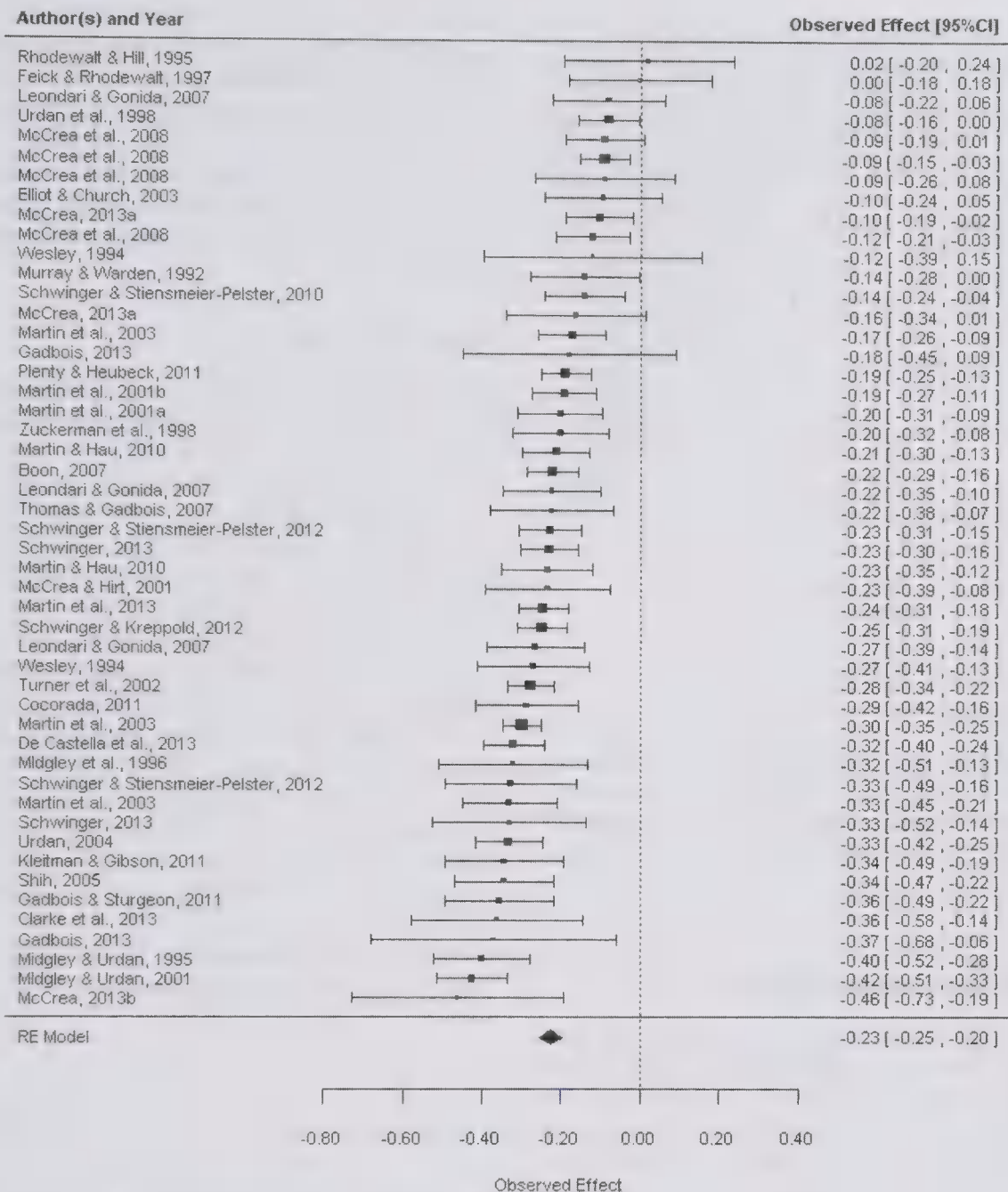


Figure 1. Forest plot of self-handicapping and achievement (correlations transformed into Fisher's Z scores). CI = confidence interval; RE = random effects.

the variable "questionnaire" emerged as a significant moderator of the mean correlation between self-handicapping and achievement ( $Q_B = 24.74, p < .01$ ). Lower correlations were found for the "SHS" ( $r = -.11$ ; 95% CI:  $-.17 \leq r \leq -.06$ ) compared to the categories "Academic SHS" ( $r = -.25$ ; 95% CI:  $-.29 \leq r \leq -.23$ ) and "MES" ( $r = -.25$ ; 95% CI:  $-.31 \leq r \leq -.20$ ). Mastery-approach goals were an additional statistically significant moderator ( $Q_B = 8.93, p < .01$ ): We found higher effect sizes in samples with medium sized mastery-approach goals ( $r = -.38$ ; 95% CI:  $-.45 \leq r \leq -.30$ ) compared to samples with high mastery-approach goals ( $r = -.25$ ; 95% CI:  $-.28 \leq r \leq -.20$ ). The reliability of the self-handicapping was an additional significant moderator (unstandardized  $b = -0.67, z = -3.0, p < .01$ ).

The effect sizes for published studies were comparable to unpublished studies, as can be seen in Table 3. Origin of the sample

was not a significant moderator as was indicated by similar correlations between self-handicapping and achievement in the United States, European, and Australian samples. The regression analysis with the continuous variable ethnicity (percentage White within a sample, just available for eight studies) was not significant as well (unstandardized  $b = -0.00, z = -0.35, p = .73$ ). With respect to gender, there were slightly higher effect sizes in samples with larger proportions of males. However, the differences were not significant. Similar findings were obtained for the moderator "time." Although the mean correlation appeared to be higher for concurrent assessments of self-handicapping and achievement, the  $Q_B$ -index did not reach significance. The mean effect sizes were also similar across several indicators of academic achievement, notwithstanding the smaller correlation between self-handicapping and test scores. Moreover, the specificity of the self-handicapping measurement, the achievement domain considered, and domain



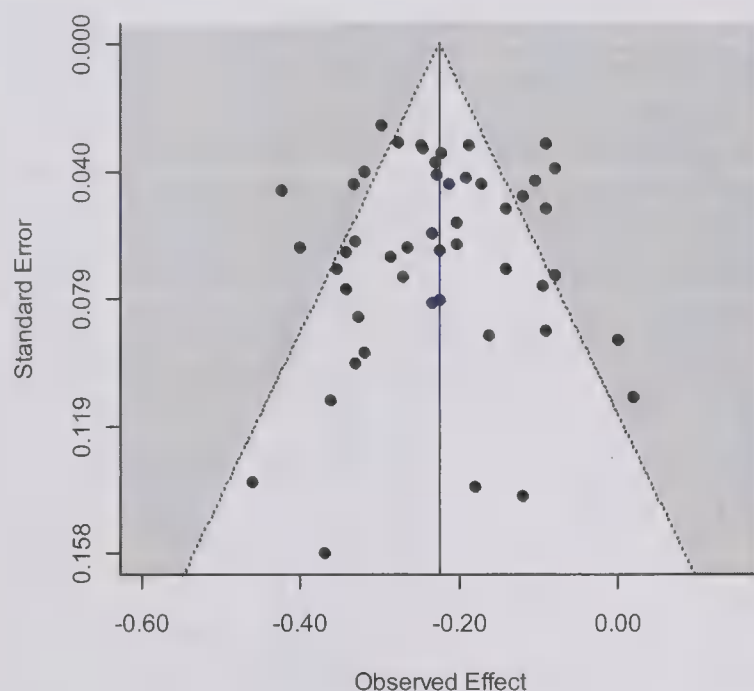


Figure 2. Funnel plot of self-handicapping and achievement.

matching did not serve as significant moderators. Performance-approach or -avoidance goals and the level of self-esteem, self-efficacy, or academic self-concept did not serve as moderator variables either. In this context, it has to be noted that the categories for these variables were comprised of only a few studies (see Table 3).

### Multiple WLS Meta-Regression Analysis

Potential meta-analytic moderators can sometimes be confounded, and this can result in spurious effects (Lipsey, 2003). With respect to the present results, we were interested in the stand-alone influences of the identified moderators. Therefore, we examined the relative effects of the three moderators school type, questionnaire, and reliability of the self-handicapping scales in a multiple WLS meta-regression analysis (see Steel & Kammeyer-Mueller, 2002; Viechtbauer, 2008). Due to the small number of existing studies, we decided to eliminate the moderator "mastery goals" from the multiple regression. Based on the results of the moderator analyses, we created two dummy variables: school type (1 = elementary and middle school students; 0 = high school and university students) and questionnaire (1 = others; 0 = SHS). We chose these categories for the regression analysis because the confidence intervals between them did not overlap. The total regression model was significant ( $Q_{model} = 39.50$ ,  $df = 3$ ,  $p < .01$ ;  $k = 36$ ) and accounted for 67.39% of the variance between the true effect sizes (i.e., of the heterogeneity). In the combined analysis, the variables school type (unstandardized  $b = -0.09$ ,  $z = -3.16$ ,  $p < .01$ ) and questionnaire (unstandardized  $b = -0.17$ ,  $z = -3.53$ ,  $p < .01$ ) retained their moderating effects.<sup>3</sup>

### Discussion

The present meta-analysis had two purposes. First, we aimed to unravel the mean correlation between self-handicapping and achievement in the academic domain. Such an analysis allowed

us to determine the relative importance of self-handicapping compared to other predictors of academic performance (Hattie, 2009). Second, we sought to identify relevant moderators of the relation between self-handicapping and achievement. Based on conceptual investigations of the two main instruments for assessing self-handicapping, we were mainly interested in whether using the ASHS (Urdan et al., 1998) versus the SHS (Jones & Rhodewalt, 1982) would yield a substantive moderator effect.

### Mean Effect Size

We identified 36 studies with 49 independent effect sizes and 25,550 participants. The mean correlation between self-handicapping and achievement was moderately negative ( $r = -.23$ ), indicating that the frequent use of self-handicapping is probably associated with poor performance. In his synthesis of over 800 meta-analyses, Hattie (2009) reported mean effect sizes for a wide range of individual and contextual predictors of academic performance. In Hattie's metric, a correlation of  $r = .20$  (which equals  $d = 0.40$ ) reflects a moderate effect size. Given that similar effect sizes were found for prominent predictors of school performance, such as the ability self-concept ( $d = 0.43$ ) or intrinsic motivation ( $d = 0.48$ ), it can be concluded that self-handicapping represents a meaningful correlate of academic achievement. Even more important, several predictors that had received great attention in educational psychology research were found to have surprisingly small effects on academic outcomes. For instance, Hattie (2009) reported effects of  $d = 0.12$  for gender,  $d = 0.29$  for homework,  $d = -0.18$  for television watching, and  $d = 0.18$  for web-based learning. These findings stress the relative importance of self-handicapping for academic achievement even further.

### Moderator Analyses

The second aim of the current meta-analysis was to analyze the relevance of potential moderator variables (e.g., self-handicapping questionnaire, school type, gender, achievement indicator, concurrent vs. prospective measurement, specificity of the self-handicapping measurement, achievement domain, domain matching). The need for moderator analyses was stressed by the heterogeneity between individual effect sizes. Four moderators were significant in the univariate analyses, namely, school type, the respective self-handicapping questionnaire, the level of mastery goals in the sample, and the reliability of the self-handicapping scale. In the multiple meta-regression analysis, school type and questionnaire remained statistically significant and were able to explain most of the existing variability between the effect sizes.

The finding regarding school type is in line with our assumption that younger students show higher relations between self-

<sup>3</sup> We report unstandardized regression coefficients, as they are more easily interpreted than standardized regression coefficients in the case of dummy-coded predictor variables. The unstandardized coefficients display the difference between the mean effect sizes of the two categories (e.g., 1 = elementary and middle school students; 0 = high school and university students) of the respective moderator variable (e.g., school type) when the other moderator (e.g., questionnaire) is statistically controlled for.

Table 3  
Results of the Moderator Analyses

Moderator	<i>k</i>	<i>ES</i>	<i>SE</i>	95% CI	<i>z</i>	<i>Q<sub>B</sub></i>
Questionnaire						24.74**
Academic SHS (Midgley & Urdan, 1995)	23	-.25**	.02	-.29, -.23	-15.57	
SHS (Jones & Rhodewalt, 1982)	11	-.11**	.03	-.17, -.06	-4.27	
Short SHS (Strube, 1986)	2	-.23**	.08	-.36, -.08	-2.92	
Mixed (Midgley & Urdan, 1995; Strube, 1986)	4	-.20**	.04	-.26, -.13	-5.50	
MES (Liem & Martin, 2012)	6	-.25**	.03	-.31, -.20	-8.49	
Others	3	-.22**	.05	-.31, -.12	-4.18	
School type						18.83**
Elementary	6	-.29**	.03	-.35, -.23	-9.20	
Middle school	4	-.34**	.04	-.41, -.25	-7.95	
High school	12	-.23**	.02	-.26, -.18	-10.14	
University	25	-.18**	.02	-.21, -.14	-9.67	
Gender						2.84
≤59% female	26	-.24**	.02	-.28, -.21	-12.65	
>59% female	17	-.19**	.02	-.24, -.15	-7.95	
Mastery-approach goals						8.93**
Medium sized/average	3	-.38**	.05	-.45, -.30	-8.78	
High	10	-.25**	.02	-.28, -.20	-11.22	
Performance-approach goals						0.29
Medium sized/average	7	-.25**	.06	-.35, -.15	-4.73	
High	1	-.35**	.17	-.60, -.02	-2.08	
Performance-avoidance goals						0.48
Medium sized/average	5	-.23**	.07	-.35, -.09	-3.32	
High	1	-.36*	.18	-.62, -.01	-2.00	
Self-efficacy						0.74
Medium sized/average	1	-0.33**	.10	-.49, -.15	-3.51	
High	8	-0.25**	.03	-.31, -.19	-7.93	
Self-esteem						0.22
Medium sized/average	2	-.29**	.09	-.44, -.13	-3.50	
High	7	-.25**	.05	-.34, -.17	-5.52	
Academic self-concept						2.89
Medium sized/average	2	-.22**	.04	-.29, -.13	-5.25	
High	4	-.28**	.02	-.32, -.25	-14.81	
Achievement						3.32
Low	2	-.23**	.08	-.36, -.07	-2.91	
Middle	33	-.24**	.02	-.27, -.19	-10.81	
High	10	-.21**	.04	-.28, -.13	-5.07	
Origin of the sample						2.66
United States	24	-.20**	.02	-.24, -.15	-9.15	
Europe	11	-.24**	.03	-.30, -.18	-7.83	
Australia	12	-.25**	.03	-.29, -.20	-9.35	
Asia	2	-.26**	.07	-.38, -.14	-4.00	
Achievement						0.03
GPA	24	-.23**	.02	-.26, -.18	-10.04	
Achievement test score	18	-.19**	.02	-.24, -.14	-7.72	
Exam grade	17	-.24**	.03	-.28, -.18	-7.96	
Semester/school report-card grade	21	-.24**	.02	-.28, -.19	-10.49	
Publication type						1.91
Published	42	-.22**	.02	-.25, -.19	-14.72	
Unpublished	7	-.23**	.04	-.32, -.14	-5.18	
Time						3.76
Concurrent	35	-.23**	.02	-.26, -.20	-14.34	
Prospective	12	-.19**	.03	-.25, -.13	-6.22	
Specificity of self-handicapping measurement						3.90
General/school	38	-.23**	.02	-.25, -.20	-14.18	
Math and science	8	-.24**	.03	-.29, -.18	-7.46	
Others	2	-.11	.06	-.23, .01	-1.76	
Achievement domain						1.75
General/school	36	-.20**	.02	-.24, -.17	-10.96	
Math and science	20	-.25**	.02	-.29, -.21	-10.56	
Other subjects	14	-.23**	.03	-.29, -.17	-6.92	
Languages	7	-.25**	.04	-.33, -.18	-6.72	
Mixed	3	-.21**	.06	-.31, -.09	-3.55	
Domain matching						

(table continues)



Table 3 (continued)

Moderator	<i>k</i>	<i>ES</i>	<i>SE</i>	95% CI	<i>z</i>	<i>Q<sub>B</sub></i>
Match	49	-.21**	.02	-.24, -.18	-11.94	
Mismatch	31	-.25**	.02	-.28, -.21	-13.74	

Note. *k* = number of effect sizes; *ES* = mean effect size; *SE* = standard error of *ES*; 95% CI = lower and upper limits of 95% confidence interval; *z* = *z* test for significance of *r*; *Q<sub>B</sub>* = homogeneity estimate; SHS = Self-Handicapping Scale; MES = Motivation and Engagement Scale; GPA = grade point average.

\* *p* < .05. \*\* *p* < .01.

handicapping and performance. This might be due to differences in grading structures (e.g., teachers of younger students might weight self-handicapping more negatively when assigning grades) and/or age-related developments (e.g., due to a poorly differentiated ability self-concept, specific failures may be interpreted to mean that the student is less capable in school in general). Moreover, the moderator questionnaire remained statistically significant in the multiple meta-regression as well. As argued in the theory section, the instruments used to assess self-handicapping differ in several ways. In fact, the SHS items are only partially in line with Urdan and Midgley's (2001) required features of a valid self-handicapping item, and the criteria that the items meet are not consistent across all SHS items. Moreover, the SHS assesses rather undifferentiated avoidance behavior, and agreement with items on the SHS can be justified by several reasons other than self-handicapping. However, self-handicapping becomes maladaptive for academic performance when the various aspects of self-handicapping all come together. Just showing a potential handicapping behavior, such as procrastination, is only one part of the self-handicapping construct. The more important parts include the a priori timing of the strategy and the reason for the behavior (e.g., procrastinating in order to have a handicap in case of failure). Because these aspects are more strongly represented by the ASHS and the respective MES subscale, it seems reasonable that the correlation between self-handicapping and achievement would be considerably higher when using these instruments. However, we cannot rule out alternative explanations of the moderating effects of the different questionnaires. With regard to the underlying specificity of self-handicapping measures, for instance, the ASHS and the MES measure self-handicapping more directly in terms of concrete behaviors, whereas the SHS rather assesses individual differences in the tendency to engage in self-handicapping behaviors. That is, the SHS operationalizes self-handicapping as a more distal construct like a broad personality trait. As a consequence, one might attribute the different effect sizes for the SHS versus other questionnaires to some kind of a bandwidth-fidelity problem (Baranik et al., 2010), resulting in the broader handicapping measure being less predictive of important outcomes. Future studies should examine this alternative interpretation in more detail.

Our findings have several important implications for both self-handicapping research and educational practice. First, one could argue that studies using the SHS are not informative when estimating the correlation between self-handicapping and achievement. It is thus crucial that self-handicapping researchers discuss the construct validity of the available questionnaires in order to gain a precise understanding of the maladaptive effects of self-handicapping in the academic domain. Second, age-related differences in the effects of self-handicapping on achievement should be

considered: The correlations were more highly negative when the students were in elementary or middle school compared to high school or university. Third, we provided several reasons for the poor construct validity of the SHS. Researchers are thus cautioned to check the face validity of self-handicapping items before using them in their studies.

Participants' gender was not found to be a significant moderator in our meta-analysis. Although it might be plausible that women are somewhat "smarter" about choosing the kind and the degree of the handicap (e.g., women may be more sensitive about not reducing their effort more than necessary), this assumption was not supported by the present data. More sophisticated investigations may clarify whether the consequences of self-handicapping are really the same for women and men.

A current meta-analysis on the relation between achievement goals and academic achievement identified the respective achievement indicator as a relevant moderator variable (Wirthwein et al., 2013). Surprisingly, our moderator analyses revealed no significant effects for different indicators of academic achievement. Although the correlation with test scores was slightly smaller compared to the other indicators, the differences were not significant. We additionally could not find a moderating effect for "achievement domain." That is, the mean effects were similar across several domains such as mathematics and languages. Given the considerable effect size when using GPA as the criterion, our findings are in line with numerous studies that have emphasized the negative long-term effects of academic self-handicapping (e.g., Martin et al., 2001a; Midgley & Urdan, 2001; Zuckerman et al., 1998). However, similar effect sizes were found for the more specific achievement criteria "exam grade" and "school report-card grade"; thus, these findings similarity contradict the often-claimed statement that singular self-handicapping events are less costly to performance. It is thus possible that students enter the "vicious cycle" of low performance and self-handicapping after some singular handicapping situations. To verify this conclusion, future studies could examine the longitudinal trajectories of students' handicapping-performance relations in more detail.

It is also important to note that self-handicapping had similar effects when using test scores versus teacher-assigned grades as the achievement criterion. This finding underlines the maladaptive impact of self-handicapping on performance. To some degree, negative effects on grades might be attributed to biased grading practices by teachers. Randall and Engelhard (2010) provided teachers with scenarios that described student ability, achievement, behavior, and effort, and asked them to assign both a numerical and letter grade. Results showed that teachers based their grades primarily on performance but to a smaller extent also on nonachievement indicators. Because teachers often interpret



self-handicapping negatively, it is possible that self-handicappers received extra deductions in marks, which would have resulted in an overestimation of the correlation between self-handicapping and achievement when using grades as the achievement criterion. However, because we failed to find a moderating effect of the achievement indicator, we conclude that the negative effects of self-handicapping are reflected not only in subjective measures of performance but also in objective ones.

The mean correlation was similar when self-handicapping was assessed with global versus domain-specific measures. Moreover, in contrast to recent meta-analyses on achievement goals (Huang, 2012; Wirthwein et al., 2013), there was no effect of domain-matching. These results suggest that self-handicapping represents a rather global construct that has similar effects across different contexts such as school domains. However, further research is needed to explore the extent to which self-handicapping in one domain (e.g., mathematics) generalizes to other school subjects (e.g., languages, natural sciences).

One of the most important questions in psychological research refers to the causality of relations, that is, the chicken–egg problem. In this meta-analysis, we were interested in the effects of self-handicapping on achievement, thereby implying that the former causally determines the latter. However, we also agree with several authors who have proposed a reciprocal effect between the two constructs (Zuckerman et al., 1998). That we did not find significant differences in concurrent versus prospective correlations may be interpreted to mean that the cross-sectional correlation coefficient might yet provide an acceptable estimation of the causal effect of self-handicapping on academic achievement. It has to be noted, however, that we did not control for previous achievement or for previous self-handicapping. Such cross-lagged analyses could yield more satisfactory conclusions about the topic of causality in future research.

A further as-yet-unmentioned reason for the substantial variability in correlation coefficients may be the distinction between behavioral and claimed self-handicapping (e.g., Arkin & Baumgardner, 1985). It is obvious that just claiming to have a handicap (such as pretending to have test anxiety or physical symptoms) is not necessarily accompanied by poor performance, whereas the active acquisition of an impediment (e.g., alcohol abuse) should be more likely to decrease one's performance. In this regard, it is interesting that the SHS includes a larger number of claimed handicapping items than the ASHS and the MES. However, the vast majority of questionnaire studies in the field have focused on the association between only a combined self-handicapping scale (i.e., one that does not differentiate between claimed and behavioral self-handicapping) and academic achievement. Hence, we were not able to take this distinction into account with regard to the mean effect size. Future research should explicitly separate these two self-handicapping strategies on questionnaires and analyze the individual consequences for achievement or achievement-related behavior (McCrea et al., 2008). Moreover, it would be interesting to examine different forms of self-handicapping separately (such as procrastinating or claiming test anxiety). Due to the small number of studies that have investigated the association between self-handicapping and achievement and the lack of a differentiated self-handicapping scale, we were not able to analyze this aspect in the current meta-analysis.

We found smaller associations between self-handicapping and achievement when the level of mastery-approach goals was high. This result should be interpreted rather cautiously because of the small number of studies considered and also because it might be attributed to the restricted variance in the group of the highly mastery-oriented students. This caveat notwithstanding, our findings suggest that a high level of mastery goals buffers the maladaptive effects of self-handicapping on achievement, an interpretation that appears to be reasonable from a conceptual perspective. At a certain point during the self-handicapping process, students realize that this behavior impedes their performance. Primarily performance-oriented students would then attribute this growing failure to internal, stable, and uncontrollable factors (e.g., low intelligence), and this attribution would probably reinforce the presumed reciprocal cycle of self-handicapping and low performance and lead them to continue handicapping. However, the additional activation of a mastery goal orientation might lead students to see the failure from a different perspective and to attribute it to controllable factors (Schwinger & Stiensmeier-Pelster, 2011). If successful, this might help students to significantly reduce the amount and/or duration of self-handicapping.

Taken together, our analyses revealed important moderators of the relationship between self-handicapping and achievement. However, the number of non-significant moderators in our study is also remarkable. The fact that self-handicapping cuts across so many very different groups and contexts is a major finding of this meta-analysis and it underlines the universality of self-handicapping effects. These results seem to indicate that self-handicapping is more trait-like than sometimes presumed or at least influenced by trait-like drivers such as fear of failure or self-esteem (Rhodewalt & Tragakis, 2002).

## Self-Handicapping Interventions

Given the considerable correlation with achievement and the substantial heterogeneity in effect sizes, it seems necessary to develop adequate educational interventions against self-handicapping. To date, specific trainings that focus explicitly on reducing self-handicapping are barely available. Kearns, Forbes, and Gardiner (2007) conducted a cognitive behavioral coaching intervention (CBC) with doctoral students in order to reduce perfectionism and self-handicapping. In a 6-week workshop series, participants learned to alter inaccurate cognitive assumptions about themselves through the use of several CBC techniques such as cognitive restructuring, thought diaries, and the normalizing of one's thoughts. Results revealed a significant decrease in self-handicapping at the follow-up assessment 4 weeks after the intervention. Martin (2005) implemented a series of workshops targeting students' motivation and engagement. Measurement involved the Motivation and Engagement Scale–High School (MES-HS; Martin, 2007) at the outset of the program, toward the end of the program, and again 6–8 weeks later. Data showed a significant reduction in self-handicapping as well as gains on key facets of students' motivation by the end of the program and also 6–8 weeks later.

Our results suggest that fostering mastery-approach goals in students might also help to reduce the amount of self-handicapping and its negative impact on academic achievement. Mastery-



oriented students believe that the self and performance are malleable (Blackwell, Trzesniewski, & Dweck, 2007) and that self-worth is not contingent on one's abilities. Consequently, they do not interpret failure as feedback concerning their self-esteem, but they view negative task experiences as opportunities for personal growth. Moreover, mastery-oriented learners tend to judge negative feedback more positively. Because they are focused on individual reference norms, they are more likely to attribute failure to modifiable and controllable factors such as low effort (Ames, 1992). In line with these considerations, Schwinger and Stiensmeier-Pelster (2011) reported that both performance-avoidance goals and low self-esteem had a lower impact on self-handicapping when mastery goals were also highly salient. Likewise, the data presented here revealed that the (additional) pursuit of mastery goals reduces the maladaptive effects of self-handicapping on academic performance. A possible interpretation may be that mastery-oriented students are more flexible in managing the duration and intensity of self-handicapping, and such an ability may help them to avoid rather extreme declines in academic performance. However, the processes behind the observed moderating effect of mastery goals remain speculative here, and further research is needed to disentangle them in more detail.

Altogether, there seem to be promising avenues by which to reduce self-handicapping, but only some of them have already been explored. These sporadic endeavors need to be extended to standardized intervention programs that are designed to prevent or minimize self-handicapping and that can be effectively applied across different forms of schooling and cultural contexts. The present meta-analysis has set the stage for understanding the importance of preventing students from becoming chronic self-handicappers.

### Limitations and Suggestions for Further Research

Some limitations need to be considered regarding the present meta-analysis. Unfortunately, due to the lack of studies that included the variables that we targeted in our moderator analyses, not all studies could be included in the moderator analyses. With respect to the selection of moderator variables, future investigations could focus on additional moderators such as the degree to which a task is challenging or difficult. In addition, it would be interesting to analyze the associations of self-handicapping not only with academic achievement indicators but also with other outcome variables such as interest, academic engagement, or the use of specific learning strategies. An important issue for further meta-analytic research on the relation between self-handicapping and achievement would be to take longitudinal cross-lagged effects into account (cf. Huang, 2011). Such examinations could shed more light on the question of which variable causes the other. Furthermore, it might be interesting to conduct a meta-analysis on the antecedents of self-handicapping so researchers can establish a rank order of the most relevant risk factors for self-handicapping. Finally, the present meta-analysis was based on questionnaire data only. Future studies may wish to focus on the large number of experimental studies in the field.

Despite the above-mentioned limitations, we believe that the present meta-analysis has provided important insights into the effects of self-handicapping on achievement in the academic domain. Considering the relevant literature up to August 2013, our

findings explicated the relative importance of self-handicapping as a correlate of academic achievement. The results should also caution researchers about which instrument to use to assess self-handicapping, as this may have an effect on the validity of the findings. Taken together, these findings have utility for practitioners, researchers, and theorists as they seek to reduce maladaptive psycho-behavioral strategies that ultimately limit students' academic potential.

### References

References marked with an asterisk indicate studies included in the meta-analysis.

- Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*, 261–271. doi:10.1037/0022-0663.84.3.261
- Arkin, R. M., & Baumgardner, A. H. (1985). Self-handicapping. In J. H. Harvey & G. W. Weary (Eds.), *Attribution: Basic issues and applications* (pp. 169–202). Orlando, FL: Academic Press.
- Baranik, L. E., Barron, K. E., & Finney, S. J. (2010). Examining specific versus general measures of achievement goals. *Human Performance, 23*, 155–172. doi:10.1080/08959281003622180
- Baumeister, R. F., Hamilton, J. C., & Tice, D. M. (1985). Public versus private expectancy of success: Confidence booster or performance pressure? *Journal of Personality and Social Psychology, 48*, 1447–1457. doi:10.1037/0022-3514.48.6.1447
- Baumgardner, A. H., Lake, E. A., & Arkin, R. M. (1985). Claiming mood as a self-handicap: The influence of spoiled and unspoiled public identities. *Personality and Social Psychology Bulletin, 11*, 349–357. doi:10.1177/0146167285114001
- Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*, 1088–1101. doi:10.2307/2533446
- Berglas, S., & Jones, E. E. (1978). Drug choice as a self-handicapping strategy in response to noncontingent success. *Journal of Personality and Social Psychology, 36*, 405–417. doi:10.1037/0022-3514.36.4.405
- Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development, 78*, 246–263. doi:10.1111/j.1467-8624.2007.00995.x
- \*Boon, H. J. (2007). Low- and high-achieving Australian secondary school students: Their parenting, motivations and academic achievement. *Australian Psychologist, 42*, 212–225. doi:10.1080/00050060701405584
- \*Clarke, I. E., MacCann, C., & Kleitmann, S. (2013). *Structure and correlates of self-handicapping in university students*. Unpublished manuscript.
- \*Cocorada, E. (2011). Academic self-handicapping and their correlates in adolescence. *Bulletin of the Transilvania University of Brasov, 53*, 57–64.
- Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. doi:10.1037/0033-2909.112.1.155
- Covington, M. V. (1992). *Making the grade: A self-worth perspective on motivation and school reform*. New York, NY: Cambridge University Press. doi:10.1017/CBO9781139173582
- Covington, M. V., & Omelich, C. L. (1979). Effort: The double-edged sword in school achievement. *Journal of Educational Psychology, 71*, 169–182. doi:10.1037/0022-0663.71.2.169
- \*De Castella, K., Byrne, D., & Covington, M. (2013). Unmotivated or motivated to fail? A cross-cultural study of achievement motivation, fear of failure, and student disengagement. *Journal of Educational Psychology, 105*, 861–880. doi:10.1037/a0032464
- Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629–634. doi:10.1136/bmj.315.7109.629

- Elliot, A. J., Chirkov, V. I., Kim, Y., & Sheldon, K. M. (2001). A cross-cultural analysis of avoidance (relative to approach) personal goals. *Psychological Science*, 12, 505–510. doi:10.1111/1467-9280.00393
- \*Elliot, A. J., & Church, M. A. (2003). A motivational analysis of defensive pessimism and self-handicapping. *Journal of Personality*, 71, 369–396. doi:10.1111/1467-6494.7103005
- Elliot, A. J., & McGregor, H. A. (2001). A  $2 \times 2$  achievement goal framework. *Journal of Personality and Social Psychology*, 80, 501–519. doi:10.1037/0022-3514.80.3.501
- \*Feick, D. L., & Rhodewalt, F. (1997). The double-edged sword of self-handicapping: Discounting, augmentation, and the protection and enhancement of self-esteem. *Motivation and Emotion*, 21, 147–163. doi:10.1023/A:1024434600296
- Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, 17, 120–128. doi:10.1037/a0024445
- \*Gadbois, S. A. (2013). *Academic self-handicapping in university students*. Unpublished manuscript.
- \*Gadbois, S. A., & Sturgeon, R. D. (2011). Academic self-handicapping: Relationships with learning specific and general self-perceptions and academic performance over time. *British Journal of Educational Psychology*, 81, 207–222. doi:10.1348/000709910X522186
- Gunderson, E. A., Ramirez, G., Levine, S. C., & Beilock, S. L. (2012). The role of parents and teachers in the development of gender-related math attitudes. *Sex Roles*, 66, 153–166. doi:10.1007/s11199-011-9996-2
- Haag, L., & Götz, T. (2012). Mathe ist schwierig und Deutsch aktuell: Vergleichende Studie zur Charakterisierung von Schulfächern aus Schülersicht [Math is difficult and German up to date: A study on the characterization of subject domains from students' perspectives]. *Psychologie in Erziehung und Unterricht*, 59, 32–46. doi:10.2378/peu2012.art03d
- Hansford, B. C., & Hattie, J. A. (1982). The relationship between self and achievement/performance measures. *Review of Educational Research*, 52, 123–142. doi:10.3102/00346543052001123
- Hattie, J. A. C. (2009). *Visible learning: A synthesis of 800+ meta-analyses on achievement*. Oxford, England: Routledge.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analyses*. Orlando, FL: Academic Press.
- Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effect models in meta-analysis. *Psychological Methods*, 3, 486–504. doi:10.1037/1082-989X.3.4.486
- Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21, 1539–1558. doi:10.1002/sim.1186
- Hirt, E. R., Deppe, R. K., & Gordon, L. J. (1991). Self-reported versus behavioral self-handicapping: Empirical evidence for a theoretical distinction. *Journal of Personality and Social Psychology*, 61, 981–991. doi:10.1037/0022-3514.61.6.981
- Hirt, E. R., & McCrea, S. M. (2009). Man smart, woman smarter? Getting to the root of gender differences in self-handicapping. *Social and Personality Psychology Compass*, 3, 260–274. doi:10.1111/j.1751-9004.2009.00176.x
- Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology*, 49, 505–528. doi:10.1016/j.jsp.2011.07.001
- Huang, C. (2012). Discriminant and criterion-related validity of achievement goals in predicting academic achievement: A meta-analysis. *Journal of Educational Psychology*, 104, 48–73. doi:10.1037/a0026223
- Jones, E. E., & Rhodewalt, F. (1982). *The Self-Handicapping Scale*. Princeton, NJ: Princeton University.
- Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *The Journal of the American Medical Association*, 282, 1054–1060. doi:10.1001/jama.282.11.1054
- Kearns, H., Forbes, A., & Gardiner, M. (2007). A cognitive behavioural coaching intervention for the treatment of perfectionism and self-handicapping in a nonclinical population. *Behaviour Change*, 24, 157–172. doi:10.1375/behc.24.3.157
- Kelley, H. H. (1971). *Attribution in social interaction*. Morristown, NJ: General Learning Press.
- \*Kleitman, S., & Gibson, J. (2011). Metacognitive beliefs, self-confidence and primary learning environment of sixth grade students. *Learning and Individual Differences*, 21, 728–735. doi:10.1016/j.lindif.2011.08.003
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174. doi:10.2307/2529310
- Leary, M. R., & Shepperd, J. A. (1986). Behavioral self-handicaps versus self-reported self-handicaps: A conceptual note. *Journal of Personality and Social Psychology*, 51, 1265–1268. doi:10.1037/0022-3514.51.6.1265
- \*Leondari, A., & Gonida, E. (2007). Predicting academic self-handicapping in different age groups: The role of personal achievement goals and social goals. *British Journal of Educational Psychology*, 77, 595–611. doi:10.1348/000709906X128396
- Liem, G. A. D., & Martin, A. J. (2012). The Motivation and Engagement Scale: Theoretical framework, psychometric properties, and applied yields. *Australian Psychologist*, 47, 3–13. doi:10.1111/j.1742-9544.2011.00049.x
- Light, R. J., Singer, J. D., & Willett, J. B. (1994). Displaying and communicating findings from a meta-analysis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 439–453). New York, NY: The Russell Sage Foundation.
- Lipsey, M. (2003). Those confounded moderators in meta-analysis: Good, bad, and ugly. *The Annals of the American Academy of Political and Social Science*, 587, 69–81. doi:10.1177/0002716202250791
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- Markus, H., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98, 224–253. doi:10.1037/0033-295X.98.2.224
- Marsh, H. W. (1992). Content specificity of relations between academic achievement and academic self-concept. *Journal of Educational Psychology*, 84, 35–42. doi:10.1037/0022-0663.84.1.35
- \*Martin, A. J. (2003). The Student Motivation Scale: Further testing of an instrument that measures school students' motivation. *Australian Journal of Education*, 47, 88–106. doi:10.1177/000494410304700107
- Martin, A. J. (2005). Exploring the effects of a youth enrichment program on academic motivation and engagement. *Social Psychology of Education*, 8, 179–206.
- Martin, A. J. (2007). Examining a multidimensional model of student motivation and engagement using a construct validation approach. *British Journal of Educational Psychology*, 77, 413–440. doi:10.1348/000709906X118036
- \*Martin, A. J., & Hau, K. (2010). Achievement motivation among Chinese and Australian school students: Assessing differences of kind and differences of degree. *International Journal of Testing*, 10, 274–294. doi:10.1080/15305058.2010.482220
- \*Martin, A. J., Marsh, H. W., & Debus, R. L. (2001a). A quadripartite need achievement representation of self-handicapping and defensive pessimism. *American Educational Research Journal*, 38, 583–610. doi:10.3102/00028312038003583
- \*Martin, A. J., Marsh, H. W., & Debus, R. L. (2001b). Self-handicapping and defensive pessimism: Exploring a model of predictors and outcomes from a self-protection perspective. *Journal of Educational Psychology*, 93, 87–102. doi:10.1037/0022-0663.93.1.87



- \*Martin, A. J., Marsh, H. W., & Debus, R. L. (2003). Self-handicapping and defensive pessimism: A model of self-protection from a longitudinal perspective. *Contemporary Educational Psychology*, 28, 1–36. doi:10.1016/S0361-476X(02)00008-5
- \*Martin, A. J., Nejad, H. G., Colmar, S., & Liem, G. A. D. (2013). Adaptability: How students' responses to uncertainty and novelty predict their academic and non-academic outcomes. *Journal of Educational Psychology*, 105, 728–746. doi:10.1037/a0032794
- \*McCrae, S. M. (2013a). [Self-esteem and performance in undergraduates]. Unpublished raw data.
- \*McCrae, S. M. (2013b). [Predicting math performance in high-school students]. Unpublished raw data.
- \*McCrea, S. M., & Hirt, E. R. (2001). The role of ability judgments in self-handicapping. *Personality and Social Psychology Bulletin*, 27, 1378–1389. doi:10.1177/01461672012710013
- \*McCrea, S. M., Hirt, E. R., Hendrix, K. L., Milner, B. J., & Steele, N. L. (2008). The Worker scale: Developing a measure to explain gender differences in behavioral self-handicapping. *Journal of Research in Personality*, 42, 949–970. doi:10.1016/j.jrp.2007.12.005
- McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, 20, 20–32. doi:10.1111/j.1745-3992.2001.tb00055.x
- \*Midgley, C., Arunkumar, R., & Urdan, T. C. (1996). "If I don't do well tomorrow, there's a reason": Predictors of adolescents' use of academic self-handicapping strategies. *Journal of Educational Psychology*, 88, 423–434. doi:10.1037/0022-0663.88.3.423
- \*Midgley, C., & Urdan, T. (1995). Predictors of middle school students' use of self-handicapping strategies. *The Journal of Early Adolescence*, 15, 389–411. doi:10.1177/0272431695015004001
- \*Midgley, C., & Urdan, T. (2001). Academic self-handicapping and achievement goals: A further examination. *Contemporary Educational Psychology*, 26, 61–75. doi:10.1006/ceps.2000.1041
- \*Murray, C. B., & Warden, M. R. (1992). Implications of self-handicapping strategies for academic achievement: A reconceptualization. *The Journal of Social Psychology*, 132, 23–37. doi:10.1080/00224545.1992.9924685
- \*Plenty, S., & Heubeck, B. G. (2011). Mathematics motivation and engagement: An independent evaluation of a complex model with Australian rural high school students. *Educational Research and Evaluation*, 17, 283–299. doi:10.1080/13803611.2011.622504
- Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education*, 26, 1372–1380. doi:10.1016/j.tate.2010.03.008
- Raudenbush, S. W. (2009). Analyzing effect sizes: Random effects models. In H. Cooper, L. V. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 295–315). New York, NY: The Russell Sage Foundation.
- R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: A qualitative study. *Teaching and Teacher Education*, 27, 472–482. doi:10.1016/j.tate.2010.09.017
- Rhodewalt, F. (1990). Self-handicappers: Individual differences in the preference for anticipatory, self-protective acts. In R. L. Higgins, C. R. Snyder, & S. Berglas (Eds.), *Self-handicapping: The paradox that isn't* (pp. 69–106). doi:10.1007/978-1-4899-0861-2\_3
- Rhodewalt, F., & Davison, J. (1986). Self-handicapping and subsequent performance: The role of outcome valence and attributional ambiguity. *Basic and Applied Social Psychology*, 7, 307–322. doi:10.1207/s15324834baspp0704\_5
- \*Rhodewalt, F., & Hill, S. K. (1995). Self-handicapping in the classroom: The effects of claimed self-handicaps on responses to academic failure. *Basic and Applied Social Psychology*, 16, 397–416. doi:10.1207/s15324834baspp1604\_1
- Rhodewalt, F., & Tragakis, M. (2002). Self-handicapping and the social self: The costs and rewards of interpersonal self-construction. In J. Forgas & K. Williams (Eds.), *The social self: Cognitive, interpersonal, and intergroup perspectives* (pp. 121–143). Philadelphia, PA: Psychology Press.
- Rhodewalt, F., & Vohs, K. D. (2005). Defensive strategies, motivation, and the self: A self-regulatory process view. In A. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 548–565). New York, NY: Guilford Press.
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). Publication bias in meta-analysis: Prevention, assessment and adjustments. doi:10.1002/0470870168
- \*Schwinger, M. (2013). Structure of academic self-handicapping—Global or domain-specific construct? *Learning and Individual Differences*, 27, 134–143. doi:10.1016/j.lindif.2013.07.009
- \*Schwinger, M., & Kreppold, M. (2012). [Self-handicapping and achievement in elementary school]. Unpublished raw data.
- \*Schwinger, M., & Stiensmeier-Pelster, J. (2010). Zusammenhänge zwischen Self-Handicapping, Lernverhalten und Leistung in der Schule [The relationship between self-handicapping, learning behavior, and achievement in school]. *Unterrichtswissenschaft*, 38, 266–283.
- Schwinger, M., & Stiensmeier-Pelster, J. (2011). Prevention of self-handicapping—The protective function of mastery goals. *Learning and Individual Differences*, 21, 699–709. doi:10.1016/j.lindif.2011.09.004
- \*Schwinger, M., & Stiensmeier-Pelster, J. (2012). Erfassung von Self-Handicapping im Lern- und Leistungsbereich: Eine deutschsprachige Adaptation der Academic Self-Handicapping Scale (ASHS-D) [Measuring self-handicapping in learning and performance domains: A German language adaptation of the Academic Self-Handicapping Scale]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 44, 68–80. doi:10.1026/0049-8637/a000061
- \*Shih, S. (2005). Taiwanese sixth graders' achievement goals and their motivation, strategy use, and grades: An examination of the multiple goal perspective. *The Elementary School Journal*, 106, 39–58. doi:10.1086/496906
- Smith, T. W., Snyder, C. R., & Handelsman, M. M. (1982). On the self-serving function of an academic wooden leg: Test-anxiety as a self-handicapping strategy. *Journal of Personality and Social Psychology*, 42, 314–321. doi:10.1037/0022-3514.42.2.314
- Smith, T. W., Snyder, C. R., & Perkins, S. C. (1983). The self-serving function of hypochondriacal complaints: Physical symptoms as self-handicapping strategies. *Journal of Personality and Social Psychology*, 44, 787–797. doi:10.1037/0022-3514.44.4.787
- Snyder, C. R., & Smith, T. W. (1982). Symptoms as self-handicapping strategies: The virtues of old wine in a new bottle. In G. Weary & H. L. Mirels (Eds.), *Integrations of clinical and social psychology* (pp. 104–127). New York, NY: Oxford University Press.
- Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin*, 133, 65–94. doi:10.1037/0033-2909.133.1.65
- Steel, P. D., & Kammeyer-Mueller, J. D. (2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology*, 87, 96–111. doi:10.1037/0021-9010.87.1.96
- Steinmayr, R., & Spinath, B. (2009). The importance of motivation as a predictor of school achievement. *Learning and Individual Differences*, 19, 80–90. doi:10.1016/j.lindif.2008.05.004
- Stipek, D., & Mac Iver, D. (1989). Developmental change in children's assessment of intellectual competence. *Child Development*, 60, 521–538. doi:10.2307/1130719
- Strube, M. J. (1986). An analysis of the Self-Handicapping Scale. *Basic and Applied Social Psychology*, 7, 211–224. doi:10.1207/s15324834baspp0703\_4

- \*Thomas, C. R., & Gadbois, S. A. (2007). Academic self-handicapping: The role of self-concept clarity and students' learning strategies. *British Journal of Educational Psychology*, 77, 101–119. doi:10.1348/000709905X79644
- Tice, D. M. (1991). Esteem protection or enhancement? Self-handicapping motives and attributions differ by trait self-esteem. *Journal of Personality and Social Psychology*, 60, 711–725. doi:10.1037/0022-3514.60.5.711
- \*Turner, J. C., Midgley, C., Meyer, D. K., Gheen, M., Anderman, E. M., Kang, Y., & Patrick, H. (2002). The classroom environment and students' reports of avoidance strategies in mathematics: A multimethod study. *Journal of Educational Psychology*, 94, 88–106. doi:10.1037/0022-0663.94.1.88
- \*Urdan, T. (2004). Predictors of academic self-handicapping and achievement: Examining achievement goals, classroom goal structures, and culture. *Journal of Educational Psychology*, 96, 251–264. doi:10.1037/0022-0663.96.2.251
- Urdan, T., & Midgley, C. (2001). Academic self-handicapping: What we know, what more there is to learn. *Educational Psychology Review*, 13, 115–138. doi:10.1023/A:1009061303214
- \*Urdan, T., Midgley, C., & Anderman, E. M. (1998). The role of classroom goal structure in students' use of self-handicapping strategies. *American Educational Research Journal*, 35, 101–122. doi:10.3102/00028312035001101
- Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A systematic review. *Educational Psychologist*, 39, 111–133. doi:10.1207/s15326985ep3902\_3
- Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, 30, 261–293. doi:10.3102/10769986030003261
- Viechtbauer, W. (2008). Analysis of moderator effects in meta-analysis. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 471–487). doi:10.4135/9781412995627.d37
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36, 1–48.
- \*Wesley, J. C. (1994). Effects of ability, high school achievement, and procrastinatory behavior on college performance. *Educational and Psychological Measurement*, 54, 404–408. doi:10.1177/0013164494054002014
- Wirthwein, L., Sparfeldt, J. R., Pinquart, M., Wegerer, J., & Steinmayr, R. (2013). Achievement goals and academic achievement: A closer look at moderating factors. *Educational Research Review*, 10, 66–89. doi:10.1016/j.edurev.2013.07.001
- \*Zuckerman, M., Kieffer, S. C., & Knee, C. R. (1998). Consequences of self-handicapping: Effects on coping, academic performance, and adjustment. *Journal of Personality and Social Psychology*, 74, 1619–1628. doi:10.1037/0022-3514.74.6.1619
- Zuckerman, M., & Tsai, F. F. (2005). Costs of self-handicapping. *Journal of Personality*, 73, 411–442. doi:10.1111/j.1467-6494.2005.00314.x

Received May 13, 2013

Revision received December 4, 2013

Accepted December 29, 2013 ■



# Capturing the Complexity: Content, Type, and Amount of Instruction and Quality of the Classroom Learning Environment Synergistically Predict Third Graders' Vocabulary and Reading Comprehension Outcomes

Carol McDonald Connor

Arizona State University and the Learning Sciences Institute

Mercedes Spencer

Florida State University and the Florida Center for Reading Research

Stephanie L. Day

Arizona State University and the Learning Sciences Institute

Sarah Giuliani

Florida State University

Sarah W. Ingebrand and Leigh McLean

Arizona State University and the Learning Sciences Institute

Frederick J. Morrison

University of Michigan

We examined classrooms as complex systems that affect students' literacy learning through interacting effects of content and amount of time individual students spent in literacy instruction along with the global quality of the classroom learning environment. We observed 27 3rd-grade classrooms serving 315 target students using 2 different observation systems. The first assessed instruction at a more micro level, specifically, the amount of time individual students spent in literacy instruction defined by the type of instruction, role of the teacher, and content. The second assessed the quality of the classroom learning environment at a more macro level, focusing on classroom organization, teacher responsiveness, and support for vocabulary and language. Results revealed that both global quality of the classroom learning environment and time individual students spent in specific types of literacy instruction covering specific content interacted to predict students' comprehension and vocabulary gains, whereas neither system alone did. These findings support a dynamic systems model of how individual children learn in the context of classroom literacy instruction and the classroom learning environment, which can help to improve observations systems, advance research, elevate teacher evaluation and professional development, and enhance student achievement.

**Keywords:** reading, classroom observation, child individual differences, intervention, language, differentiated instruction

Reading comprehension and vocabulary have been identified as strong predictors of future academic success (National Institute of Child Health and Human Development [NICHD], 2000) as well as of overall school and life outcomes (Beck, McKeown, & Kucan, 2002). Yet, by the end of fourth grade, only about 34% of U.S. students are reading and comprehending proficiently (National Center for Education Statistics, 2013). Accumulating research

points to the importance of classroom literacy instruction and the opportunities to learn that students receive in the early grades (Connor et al., 2013; NICHD, 2000; Pianta, Belsky, Houts, & Morrison, 2007; Snow, 2001; Tuyay, Jennings, & Dixon, 1995). Understanding the classroom learning environment is important, and finding ways to elucidate the active ingredients of this environment that predict student outcomes are essential but challenging.

This article was published Online First February 24, 2014.

Carol McDonald Connor, Department of Psychology and the Learning Sciences Institute, Arizona State University; Mercedes Spencer, Department of Psychology and the Florida Center for Reading Research, Florida State University; Stephanie L. Day, Department of Psychology and the Learning Sciences Institute, Arizona State University; Sarah Giuliani, Department of Communication Sciences, Florida State University; Sarah W. Ingebrand and Leigh McLean, Department of Psychology and the Learning Sciences Institute, Arizona State University; Frederick J. Morrison, Department of Psychology, University of Michigan.

This study was funded by U.S. Department of Education, Institute of Education Sciences Grants R305H04013 and R305B070074, "Child by

Instruction Interactions: Effects of Individualizing Instruction," and by Eunice Kennedy Shriver National Institute of Child Health and Human Development Grants R01HD48539, R21HD062834, and P50 HD052120. The opinions expressed are ours and do not represent views of the funding agencies. We thank Elizabeth Crowe, Stephanie Day, Jennifer Dombek, and the ISI Project team members for their hard work providing professional development, collecting data, and coding video. We thank the children, parents, teachers, and school administrators without whom this research would not have been possible.

Correspondence concerning this article should be addressed to Carol McDonald Connor, Learning Sciences Institute, Arizona State University, P.O. Box 872111, Tempe, AZ 85287-2111. E-mail: Carol.Connor@asu.edu

In this study, we used a dynamic systems framework (Yoshikawa & Hsueh, 2001), which holds that there are multiple sources of influence on children's learning (Bronfenbrenner & Morris, 2006), including the instruction they receive, how this instruction is delivered (Connor, Piasta, et al., 2009; Reis, McCoach, Little, Muller, & Kaniskan, 2011), the general climate of the classroom (Rimm-Kaufman, La Paro, Downer, & Pianta, 2005), teacher characteristics (Raver, Blair, & Li-Grining, 2011), and students themselves (Connor & Morrison, 2012; Justice, Petscher, Schatschneider, & Mashburn, 2011). Further, these sources of influence interact in different ways, with some seemingly important factors (e.g., teacher education) having relatively small effects on students' reading development (Goldhaber & Anthony, 2003) and other factors (e.g., content and minutes of instruction) having large effects (Connor, Morrison, Schatschneider, et al., 2011). High-quality literacy instruction should provide students with individualized opportunities to learn that, in turn, influence their reading comprehension and language development (Beck et al., 2002; Beck, Perfetti, & McKeown, 1982; Connor et al., 2013; Snow, 2001). Thus, there is an increasing policy and research focus on how to measure classroom instruction in ways that validly and robustly predict gains in students' literacy and vocabulary skills (see Crawford, Zucker, Williams, Bhavsar, & Landry, in press; Kane, Staiger, & McCaffrey, 2012; Ramey & Ramey, 2006; Reddy, Fabiano, Dudek, & Hsu, in press; Whitehurst et al., 1988). The aim of this study was to systematically investigate the classroom-learning environment as a dynamic system, identify major dimensions of classroom instruction—at both the individual student level and the global classroom level—that may influence students' literacy achievement, and determine how these dimensions might work together synergistically to support (or fail to support) opportunities for learning that result in gains in third graders' vocabulary and reading comprehension.

### Classroom Observation Systems

Teacher value-added scores have revealed that there is measurable variability in the effectiveness of teaching, which has direct implications for students' success or failure (Konstantopoulos & Chung, 2011). However, value-added scores do not reveal what is going on in the classroom and the characteristics of the environment that explain the variability in teachers' value-added scores. The development of rigorous classroom observation systems that are reliable and have good predictive validity are important because they help to open up the black box of classroom instruction, so to speak, and begin to move us toward what has been described as "shared instructional regimes" (Raudenbush, 2009). Raudenbush (2009) described historical and recent theories of teaching as "privatized idiosyncratic practice" (p. 172) whereby teachers close their classroom doors and teach in the ways they believe to be best and where the ideal teacher develops his or her own curriculum. The "idiosyncratic" practice of teachers who have a good grasp of the current research, who have expert and specialized knowledge of their content area, and who understand how to use research evidence to inform their practice can be highly effective. However, the privatized idiosyncratic practice of some teachers may be highly ineffective (Piasta, Connor, Fishman, & Morrison, 2009), particularly for children from low-socioeconomic status families whose home learning environment and access to resources is

limited and who are more reliant on the instruction they receive at school. Research-based observation tools allow us to illustrate what effective expert practice in the classroom actually looks like so that it can be shared among a community of professionals—both educators and researchers—to improve teaching.

There are several well-documented observation systems in use with new systems being developed (Connor, 2013a). These classroom observation systems provide important insights, and most of them explain at least modest amounts of the variance in students' literacy learning. For example, Kane and colleagues (2012) tested several observation instruments, including the Framework for Teaching (Danielson, 2007), CLASS (Pianta, La Paro, & Hamre, 2008), Protocol for Language Arts Teaching (Grossman et al., 2010), Mathematical Quality of Instruction (Hill, Ball, Bass, & Schilling, 2006), and UTeach Teacher Observation Protocol (2009). Results revealed that although none of the systems designed to assess English/Language Arts instruction correlated with teacher value-added scores computed using state-mandated assessments of English/Language Arts, they were mildly to moderately positively correlated with teacher value-added scores computed using the SAT-9 reading assessment.

### Classroom Observations in the Present Study

We used two different observation coding systems to test the dynamic systems model of instruction in the present study: the quality of the classroom learning environment (CLE) and Individualizing Student Instruction (ISI)/Pathways-observation system (ISI/Pathways; Connor, Morrison, et al., 2009). The first was designed to capture the global quality of the CLE using a rubric that captured elements of the CLE that are generally predictive of student outcomes. The second, ISI/Pathways, was designed to record the amount of time individual students spent in various types of literacy instruction, the content of this instruction, the role of the teacher, and the context (e.g., whole class, small group) in which instruction was provided. We conjectured, following the dynamic systems model, that classroom opportunities to learn would operate at both student and classroom levels and that the two systems together might better elucidate the complexities of the classroom and effective learning opportunities afforded to students than either system alone. We describe each below.

### Quality of the CLE Rating Scale

The CLE rating scale (see Appendix A) was designed to rate the classroom on three dimensions: Teacher Warmth, Responsiveness and Discipline; Classroom Organization; and Teacher Support for Vocabulary and Language Development—with one rating for each scale for the entire observation of the literacy block. *Teacher warmth, responsiveness, and discipline* were defined as teachers' regard for their students, the overall emotional climate of the classroom, as well as the way in which they responded to students, particularly with regard to how they responded to student misbehavior and disruptions (Pianta, La Paro, Payne, Cox, & Bradley, 2002). Examples of teacher warmth and responsiveness include being supportive of students, providing positive feedback, clearly communicating what is expected of students, and providing discipline in a positive and supportive way (Rimm-Kaufman et al., 2005). The kinds of discussions and types of questions used, for



example, coaching versus telling (Taylor & Pearson, 2002), were measured indirectly through this dimension. Research has shown that students whose teachers were more warm and responsive achieved greater gains in reading skills, including vocabulary, by the end of first grade (Connor, Son, Hindman, & Morrison, 2005).

*Classroom organization* is defined as the degree to which the teacher takes time to give students thorough directions for upcoming activities, has clear rules for behavior, and has established routines that optimize student learning time (Wharton-McDonald, Pressley, & Hampston, 1998). When teachers have strong orienting and organizational skills, they are better able to create an efficient and productive CLE. It has been found that teachers who implement rules and effectively establish routines are less likely to have difficulties with classroom management (Borko & Niles, 1987; Cameron, Connor, Morrison, & Jewkes, 2008).

According to Beck and colleagues (2002), *teacher support for vocabulary and language development* should be “robust,” meaning that instruction should include activities beyond those that encourage rote memorization of words and their definitions and, instead, involve rich contexts that extend beyond the classroom. Such support has the potential to improve language skills overall given that processes where vocabulary knowledge is highly used (e.g., during reading comprehension) require skills above and beyond knowing the definitions of words. Therefore, instructional techniques that take into consideration that vocabulary knowledge is part of students’ background knowledge (Stahl, 1999), rather than a singular component, are more likely to be effective. Such techniques encourage students to actively use and think about word meanings and create word associations in multiple contexts. Accumulating evidence further highlights the importance of supporting student language development because vocabulary (and oral language skills in general) are highly predictive of students’ reading comprehension (Biemiller & Boote, 2006; Cain, Oakhill,

& Lemmon, 2004) and, at the most basic level, allow students to understand grade-level texts. Yet, despite the known importance of vocabulary and language skills to later reading abilities and subsequent academic success, it has been shown that vocabulary instruction is often missing from language arts/literacy classrooms (see Cassidy & Cassidy, 2005/2006; Rupley, Logan, & Nichols, 1998).

**The ISI/Pathways Classroom Observation System (ISI/Pathways)**

The ISI/Pathways system (Connor, Morrison, et al., 2009) measures the amount of time (minutes;seconds) individual students within a classroom spend in literacy instruction activities across three dimensions: content of instruction; context; and the role of the teacher and student in the learning activity. The *content of instruction dimension* (see Table 1 and Appendix B) captures the specific topic of the literacy instruction that individual students are receiving (e.g., comprehension, vocabulary, text reading). We also coded noninstructional activities, which included off-task or disruptive behaviors, transitions between activities, or time spent when the teacher was giving directions for upcoming activities. The *context dimension* captures the student-grouping arrangement and includes whole class, small group, or individual instruction. *Management* captures who is controlling the students’ attention during an activity: the teacher and student working together (teacher–child managed), peer-managed (students working with each other), or child–self-managed (student managing his or her own attention). These dimensions operate simultaneously (see Table 1) to describe instructional and noninstructional activities observed during reading instruction. For example, the teacher and students discussing a book they just read together would be coded

Table 1  
*Dimensions of Instruction (Context, Grouping, and Management) and Content Areas Associated With Code- and Meaning-Focused Types of Instruction*

Variable	Code-focused	Meaning-focused
Teacher/child-managed, Whole class	The teacher is teaching the class how to decode multisyllabic words by writing them on the white board and then demonstrating various strategies, such as looking for prefixes and suffixes.*	The teacher is reading <i>A Single Shard</i> to the class. She stops every so often to ask the students questions.
Teacher/child-managed, Small group	The teacher is working with a small group of children on spelling (i.e., encoding) strategies.	The teacher and a small group of students are discussing <i>Mr. Poppers’ Penguins</i> and how it is similar and different from <i>Charlotte’s Web</i> .
Child/peer-managed, Small group and individual self-managed	Students are working together in pairs to complete a worksheet on dividing multisyllabic words into syllables.	Students are working individually to revise an argumentative essay using feedback from their peers.
Content areas (in the coding system)	Phonological Awareness Morphological Awareness* Word Identification/Decoding Word Identification/Encoding Grapheme/Phoneme correspondence Fluency*	Print and Text Concepts Oral Language Print Vocabulary Listening and Reading Comprehension Text Reading Writing

\* It can be argued that morphological awareness (Carlisle, 2000) and fluency (Therrien, 2004) might also be considered meaning-focused activities. In our theory of literacy instruction, code-focused activities represent the more automatic processes, whereas meaning-focused activities require the integration of the more automatic processes with active construction of meaning of connected text. Hence, code-focused activities are more likely to directly affect aspects of reading related to skill, whereas meaning-focused activities are more likely to directly contribute to aspects of comprehension and reading for understanding.

as comprehension (meaning-focused) teacher/child managed, whole class activity that lasted for 11 min.

A key characteristic of ISI/Pathways is that the measurement of instructional time and content is assessed for individual students in the classroom (Connor, Morrison, et al., 2009). Hence, the system is able to capture the learning opportunities afforded to each student, for example, recording that Student A was reading with the teacher while, at the same time, Student B was off task and not redirected. A global classroom-level system would likely capture Student A's instructional opportunities but not Student B's. The more precise measure of each individual student's learning opportunities has been used to identify instructional practices as well as Child Characteristic  $\times$  Instruction interaction effects on students' reading achievement (Connor, Morrison, Fishman, et al., 2011; Connor, Morrison, Schatschneider, et al., 2011; Connor, Piasta, et al., 2009). Across studies, amounts, content, context, and types of instruction measured by the ISI/Pathways observation system predicted student literacy achievement (Connor, Morrison, et al., 2009), particularly the difference between observed and recommended individualized types/content and amounts of instruction (Connor, Piasta, et al., 2009). The closer the observed amount matched the recommended amount, the greater the students' literacy gains were.

We posed the following research question:

*How does combining measures of the duration of different types of literacy instruction and content for individual students with a more global measure of the CLE synergistically affect students' reading comprehension and vocabulary outcomes?* Using our dynamic systems model of the classroom, we hypothesized that neither the quality of the CLE nor the amount of time individual students spent in different types/content of literacy instruction (ISI/Pathways) would be strong independent predictors of vocabulary and comprehension outcomes for third-grade students. Rather, we hypothesized that there would be interaction effects involving both systems that would significantly and positively predict third graders' language and comprehension gains. Such interaction effects would, hypothetically, better capture the complexity of classroom instruction and the learning environment.

## Method

### Participants

The participants included third-grade teachers ( $n = 27$ , 13 in the individualized reading group and 14 in the vocabulary control group) and their students ( $n = 315$ ) in seven schools who were participating in a randomized controlled study evaluating the efficacy of individualized reading instruction from first through third grade (Connor, Morrison, Fishman, et al., 2011). We selected third grade because comprehension of text becomes increasingly important (Gottardo, Stanovich, & Siegel, 1996; Reynolds, Magnuson, & Ou, 2010) as children move from *learning to read* to *reading to learn* (Chall, 1967). The schools were located in an economically and ethnically diverse school system in north Florida.

**Teachers.** All teachers completed the study with the exception of three teachers who were not present during the last month of the study; however, results for these teachers and their students were used in the analysis because observations were completed before they left, and all of their participating students were assessed. All

of the teachers met the state certification requirements and had at least a bachelor's degree related to an educational field. Seven of the teachers had certifications or degrees beyond a bachelor's degree. Teachers' classroom teaching experience ranged from 0 to 30 years, with a mean of 10.9 years of experience.

Teachers in both conditions participated in half-day workshops in the fall and again in January for either literacy or vocabulary. They also participated in monthly meetings. Teachers in the reading intervention also received biweekly classroom-based support (not on the day observed). In total, teachers in the vocabulary condition received about 12 hr of professional development, and teachers in the individualized reading intervention received between 18 and 20 hr (some needed more help with the technology). Professional development for the individualized reading intervention helped teachers learn how to individualize student instruction and how to be better organized. Professional development for the vocabulary intervention focused on vocabulary teaching methods described in Beck et al. (2002). Results of this study revealed that students whose teachers were in the individualized reading intervention group demonstrated significantly greater gains in reading comprehension than did the students whose teachers were in the vocabulary intervention group. Results are fully described in Connor et al., 2011.

**Students.** Schoolwide percentages of students qualifying for free and reduced lunch (FARL) programs ranged from 92% (high poverty) to 4% (affluent). All schools used the Open Court Reading Curriculum and had a 90-min uninterrupted block of time devoted to reading instruction. All observations were conducted during this literacy block.

Student participant demographics were collected through parent reports and school records and were as follows: 36% of the students were White, 51% were African American/Black, 3% were Hispanic, 3% were Asian/Asian American, 3% were multiracial, and the remaining 4% indicated other ethnic groups. Forty-seven percent qualified for FARL. Approximately 12% qualified for special education services. A subset of students from each classroom was randomly selected to be the focus of observation coding using the following procedure: Because this was a first- through third-grade longitudinal study (Connor et al., 2013), third graders who were in the first- and second-grade studies were automatically selected as target students. We then randomly selected from among their classmates to bring the total number of target students to a minimum of eight per classroom after rank ordering and randomly selecting within terciles so that we had a distribution of reading skill level. This provided the final sample of 315 students. On average, there were 11 target students per classroom, and this ranged from six to 19. Three classrooms had six target students; two had seven students; all others had eight or more target students. Comparisons of this subsample with the entire sample revealed no significant differences on any of the measures of interest.

### Observation of Instruction and CLE

Again, in this study, we used two different observation systems—the ISI/Pathways system, which measured the amounts and types/content of instruction students received, and the CLE quality rubric, which captured the quality of the classroom learning environment. Both systems used the same videotaped classroom ob-



servations. Classrooms were videotaped three times, once during the fall, once in winter, and once in the early spring of the academic year. This video footage of the 90-min block of time devoted to literacy instruction was captured using two digital camcorders with wide-angle lenses. Cameras were not focused on individual students per se. Rather, cameras were positioned at opposite sides of the classroom to capture as much of the class as possible. However, during small group instruction, one camera was focused on the teacher's small group, whereas the second camera captured students working independently and the other small groups. While video recording, trained research assistants kept detailed field notes regarding the activities and materials used, including careful descriptions of target students and activities of students who might be off camera (Bogdan & Biklen, 1998). These notes were used in conjunction with video footage during coding and provided information for coders about students or activities that could not easily be seen on the videos. Observations were scheduled at the teachers' convenience, so the assumption was that the instruction was of the highest quality the teacher could provide.

**Quality of the CLE.** The quality of the CLE was assessed using a detailed rubric/rating scale (see Appendix A). The rating scale ranged from 1 (*low*) to 6 (*high*) and examined three global classroom-level dimensions—organization, support for vocabulary and language, and teacher responsiveness—with priority given to specific aspects of the CLE that were the focus of the professional development provided and that previous research has associated with more effective instruction (i.e., higher quality; Brophy, 1979; Cameron, Connor, & Morrison, 2005; Pianta et al., 2002; Snow, Burns, & Griffin, 1998; Taylor, Pearson, Clark, & Walpole, 2000; Wharton-McDonald et al., 1998). Trained research assistants coded all three videotaped observations using the CLE rubric. Highly trained research assistants who were blind to the teachers' treatment assignment coded the video footage. Sufficient interrater reliability on fall observations, based on Landis and Koch's (1977) criteria, was reached prior to coding the winter observation (Cohen's  $\kappa = 0.73$ ). Approximately 10% of coded winter and spring footage was randomly selected and recoded, and interrater reliability was maintained (Cohen's  $\kappa = 0.73$ ). The winter observation CLE score was used in this study because teaching tends to be more consistent during the winter months (Hamre, Pianta, Downer, & Mashburn, 2007) than in the earlier months, when routines are getting established and teachers are just getting to know their students. Two teachers' scores were based on the spring observation because a student teacher, and not the primary teacher, was teaching during the winter observation. Because scores on the three scales were moderately correlated ( $r = .59-.60$ ) and combining the three scores improved internal reliability, the scores were summed to provide a total CLE score.

**ISI/Pathways observation system.** As noted previously, the ISI/Pathways system was designed to document instruction across three dimensions: (a) the *content* of the literacy instruction (e.g., comprehension, oral language); (b) the *context* (i.e., small group, individual, or whole class); (c) the *extent to which the teacher was interacting with students (management)*, which included teacher/child-managed instruction (teacher and students working together) or child-managed instruction (students working with each other or independently) (Connor, Morrison, et al., 2009). Using Noldus Observer Video-Pro software (XT version 8.0; Noldus, Trienes, Hendriksen, Jansen, & Jansen, 2000), instructional activities ob-

served during the 90-min literacy block, which lasted 15 s or longer, were coded for each target student. Noninstructional practices (including transitions) were also coded. An excerpt from the ISI/Pathways Coding Manual (Connor, Morrison, et al., 2009) is presented in Appendix B.

Trained research assistants coded each of the videos using the Noldus Observer XT software. The training process was extensive, typically lasting 4–6 weeks until coders achieved adequate interrater reliability (Cohen's  $\kappa > 0.7$ ) with a master coder. Questions about coding were discussed at biweekly coding meetings until consensus was achieved. Random selection and analysis of approximately 10% of the videos revealed good ongoing interrater reliability among the coders (mean Cohen's  $\kappa = 0.72$ ; Landis & Koch, 1977). The lengths of videos varied slightly depending on the season. Fall mean observation length was 85 min ( $SD = 30$ ), 73 min in the winter ( $SD = 35$ ), and 79 min in the spring ( $SD = 23$ ).

The observation system provided a detailed description of over 200 instructional variables. For this study, we separated type of instruction into code-focused and meaning-focused instruction (see Table 1). Code-focused instruction consisted of five distinct types: phonological awareness, morpheme awareness, word decoding, word encoding, and fluency. Meaning-focused instruction consisted of six distinct types: print and text concepts, oral language and oral vocabulary, print vocabulary, listening and reading comprehension, text reading, and writing. These distinctions were made using our theoretical framework that the largely unconscious and more automatic processes involved in reading, including subsentence processes (e.g., morphological awareness), and those that supported fluency (i.e., automaticity) were considered code-focused, whereas more reflective and text-level processes were meaning-focused (Connor, 2013b). The case can be made that morphological awareness should be considered a meaning-focused activity (Carlisle, 2000). Unfortunately, too little morphological awareness instruction was observed to test this alternative. For each student, the time (in seconds) spent in particular types of instruction (e.g., teacher/child-managed small group meaning-focused) were computed for fall, winter, and spring and then aggregated by taking the mean amount for each student.

## Student Assessments

Trained research assistants assessed students' language and literacy skills in the fall and again in the spring. Two measures of comprehension and two of vocabulary were used in this study. Comprehension was assessed using the Woodcock-Johnson-III Passage Comprehension subtest (WJ-III; Woodcock, McGrew, & Mather, 2001) and Level 3 Gates-MacGinitie Reading Tests (GMRTs; MacGinitie & MacGinitie, 2006) Reading Comprehension subtest. Vocabulary was assessed using the WJ-III Picture Vocabulary subtest and the GMRT Reading Vocabulary subtest. Alternate versions of the assessments were administered in the fall and spring. Scores were provided to teachers and parents.

The WJ-III Passage Comprehension uses a cloze procedure in which students read a sentence or short passage and supply the missing word. The Picture Vocabulary task asks students to name increasingly unfamiliar pictures. These subtests are administered individually and have demonstrated alpha reliability estimates of .88 for passage comprehension and .81 for picture vocabulary. W

scores were used in the analyses because they are on an equal-interval scale, similar to the Rasch score (Rasch, 2001).

In the GMRT Reading Comprehension assessment, students read passages of varying length and complexity based on excerpts from narrative and expository texts commonly used in schools. Students are then asked to answer multiple-choice questions, including some that require fairly high-level inferencing. As the student progresses through the test, the text becomes more difficult and the questions demand more inferencing. The GMRT multiple-choice vocabulary assessment requires students to read and then choose the correct meaning of an underlined word within a short phrase from four possible answers. Alternate form reliability for the GMRT has been reported to range from .74 to .92, and test-retest reliability has been reported as ranging from .88 to .92. Extended scale scores were used for data analyses, which have an equal interval scale.

### Analytic Strategies

**Structural equation modeling.** We used structural equation modeling to examine the constructs of language and comprehension, and we hypothesized that the language and literacy measures would comprise one latent variable (e.g., Mehta, Foorman, Branum-Martin, & Taylor, 2005), keeping in mind that the GMRT vocabulary assessment required the students to read. As anticipated, the four measures were correlated (see Table 2). We used confirmatory factor analyses (AMOS version 21) to compare one- and two-factor models (Kline, 1998). We provide results for the spring models in detail. Results were highly similar for the fall models. The two-factor model (comprehension and vocabulary) provided only moderately acceptable fit (Tucker-Lewis Index [TLI] = .939; comparative fit index [CFI] = .994; root-mean-square error of approximation [RMSEA] = .110; Akaike's information criterion [AIC] = 32.401), whereas the one-factor model provided a superior fit (TLI = .969; CFI = .994; RMSEA = .079; AIC = 31.525). We created fall and spring factor scores using principal component factor analysis, which explained 72.8% of the variance in the fall scores and 71.3% of the variance in the spring

scores. Loadings are provided in Table 2. The factor variable, Vocabulary/Comprehension (VocComp;  $z$  score with a mean of 0 and an  $SD$  of 1), was used in the analyses. Such factor or latent variable scores have the advantage of reducing measurement error and better capturing the complex construct of interest (Keenan, Betjemann, & Olson, 2008; Kline, 1998; Mehta et al., 2005; Vellutino, Tunmer, Jaccard, Chen, & Scanlon, 2004).

**Hierarchical linear modeling (HLM).** Due to the nested structure of the data, students nested within classrooms and schools, HLM (Raudenbush & Bryk, 2002) was used to answer our research question. HLM analyses accounted for shared variance within classrooms and schools, resulting in more accurate effect sizes and noninflated standard errors (Raudenbush & Bryk, 2002). Model specification occurred in several steps. Initially, an unconditional model was created. Variance at the classroom level was divided by total variance (the summation of student- and classroom-level variance) to obtain the intraclass correlations (ICCs). ICC values represent the proportion of variance falling between classrooms. A three-level model with students nested within classrooms and classrooms nested within schools was created. This model indicated no significant between-school variance. Therefore, a simpler two-level model with the spring VocComp factor score as the outcome ( $Y_{ij}$ ) was created. Starting with the unconditional model, we first created a model with only the ISI/Pathways variables entered at the student level. Next, we created a model with only CLE quality entered at the classroom level; we then created a combined model and tested for cross-level interaction effects (ISI/Pathways [student]  $\times$  CLE Quality [classroom]). We also tested for students' fall VocComp Score  $\times$  Instruction interactions (models are available upon request).

### Results

Students in this sample entered third grade with vocabulary skills generally in line with grade-level expectations based on standard scores ( $M = 100$ ,  $SD = 15$ ), and percentile ranks ( $M = 50$ ), which control for age. For example, their fall mean WJ Picture Vocabulary score was 98.92 and their spring mean score was 99.21

Table 2

*Correlations, Means, Standard Deviations, Standard Scores, Percentile Rank, and Factor Loadings of the Reading Comprehension and Vocabulary Measures Administered During the Fall and Spring of the Academic Year*

Variable	1	2	3	4	5	6	7	8
1. Fall WJ-PC	—							
2. Fall WJ-PV	.579**	—						
3. Fall GM-C	.627**	.510**	—					
4. Fall GM-V	.705**	.669**	.749**	—				
5. Spring WJ-PC	.697**	.517**	.602**	.663**	—			
6. Spring WJ-PV	.503**	.821**	.451**	.633**	.523**	—		
7. Spring GM-C	.622**	.534**	.780**	.740**	.624**	.504**	—	
8. Spring GM-V	.701**	.674**	.707**	.847**	.670**	.610**	.760**	—
<i>M</i>	94.75 <sup>a</sup>	98.92 <sup>a</sup>	51.81 <sup>b</sup>	55.04 <sup>b</sup>	95.96 <sup>a</sup>	99.21 <sup>a</sup>	52.07 <sup>b</sup>	59.84 <sup>b</sup>
<i>SD</i>	10.31 <sup>a</sup>	10.82 <sup>a</sup>	28.04 <sup>b</sup>	27.26 <sup>b</sup>	10.43 <sup>a</sup>	10.79 <sup>a</sup>	29.79 <sup>b</sup>	25.52 <sup>b</sup>
Maximum	124 <sup>a</sup>	127 <sup>a</sup>	99 <sup>b</sup>	99 <sup>b</sup>	121 <sup>a</sup>	139 <sup>a</sup>	99 <sup>b</sup>	99 <sup>b</sup>
Minimum	48 <sup>a</sup>	68 <sup>a</sup>	1 <sup>b</sup>	1 <sup>b</sup>	54 <sup>a</sup>	69 <sup>a</sup>	1 <sup>b</sup>	2 <sup>b</sup>
Factor loadings	.85	.798	.845	.917	.835	.861	.768	.908

Note. WJ = Woodcock-Johnson; PC = passage comprehension; PV = print vocabulary; GM = Gates-MacGinitie; C = comprehension; V = vocabulary.

<sup>a</sup> Standard scores. <sup>b</sup> Percentile rank.

\*\*  $p < .001$ .



(see Table 2), suggesting grade-/age-appropriate gains from fall to spring, on average. On the basis of GMRT results, students were achieving slightly above grade expectations for both reading comprehension and reading vocabulary, with percentiles in the fall of 51.81 and 55.04, respectively, and spring scores of 52.07 and 59.84, respectively. Only on the WJ Passage Comprehension measure did students generally score below test expectations, with standard scores of 94.75 in the fall and 95.96 in the spring. Notably, students' standard scores and percentile ranks were the same or higher in the spring compared with the fall, suggesting generally grade-appropriate gains in reading and vocabulary.

Overall, we observed high-quality but variable CLE in these third-grade classrooms, with total scores ranging from 5 to 17, where a perfect score was 18 ( $M = 12.58$ ,  $SD = 3.14$ ). When considering the individual scales, teachers generally scored highest on the Orienting/Planning scale ( $M = 4.73$ ,  $SD = 1.09$ ) and lowest on the Support for Language scale ( $M = 3.66$ ,  $SD = 1.21$ ). They were rated a mean of 4.30 ( $SD = 1.34$ ) on the Responsiveness/Discipline scale.

An examination of the kinds of literacy activities that occurred during reading comprehension and vocabulary instruction using the ISI/Pathways observation system revealed variability in overall amounts and types of instruction (see Table 3, Figure 1, and Appendix C). Beginning first with oral language and print vocabulary instruction, on average, students received about 2 min per day of small group teacher/child-managed instruction and about 3 min per day of small group child-managed instruction. Of this time, most was spent using vocabulary (Vocabulary Use, see Appendix B) and defining words with many of the child-managed activities using workbooks. See Figure 1 for graphs showing the

amounts per day (seconds) of the various types of vocabulary activities observed. Examining teacher/child-managed small group instruction in listening and reading comprehension revealed that students spent approximately 4 min per day in these activities, with most of the time spent with the teacher asking questions and students responding (2.2 min). Generally, more than 7 min per day were spent in whole-class teacher/child-managed vocabulary and reading comprehension activities, with most of that time spent in question-and-answer time (3.1 min). In all cases, the ranges were large, with some students receiving little vocabulary and comprehension instruction and some receiving much more.

Using these data, we created eight different variables of amounts and types/content of instruction (see Tables 1 and 3) that capture the entire duration of meaningful instruction that occurred during the literacy block: teacher/child-managed whole-class meaning-focused instruction; teacher/child-managed small group meaning-focused instruction; teacher/child-managed whole-class code-focused instruction; teacher/child-managed small group code-focused instruction; child/peer-managed whole-class meaning-focused; child/peer-managed small group meaning-focused; child/peer-managed whole-class code-focused; and child/peer-managed small group code-focused instruction. To create these variables, the data were first exported from Observer Pro and then cleaned and examined in SPSS (Version 20). Because there were multiple tapes that were coded for one observation, these were aggregated for each student by summing the seconds for each activity within a type of instruction (e.g., comprehension, print vocabulary), providing a total amount of instruction for each student.

The amounts of each type of instruction were then combined on the basis of our theory of reading instruction (Connor, Morrison, & Katch, 2004; Connor, Morrison, & Petrella, 2004). Meaning-focused instruction was composed of all instructional activities that might be expected to *explicitly* support students' language and comprehension skill gains, including text reading, writing, oral language, listening and reading comprehension, vocabulary, and other meaning-focused types of instruction (see Table 1). There were also activities that might be expected to support language and comprehension more *implicitly* through instruction in how to decode unfamiliar words and building automaticity, which might be considered more code-focused types of instruction. Descriptive statistics for each type of instruction are provided in Table 3. Of note, most of the time was spent in teacher/child-managed whole-class meaning-focused instruction (12.6 min per day), followed by teacher/child-managed small group meaning-focused (9.32 min per day). Children spent about 11 min per day working with peers or individually on meaning-focused activities. Less time was spent in code-focused activities, about 6 min per day compared with a total of about 19 min in meaning-focused activities.

How does combining measures of the duration of different types of instruction and content for individual students with a more global measure of the CLE synergistically affect students' reading comprehension and vocabulary outcomes?

As noted in the Method section, we ran three different models, with the spring VocComp score as the outcome and controlling for fall VocComp scores. The ICC for the unconditional model, which is the between-classroom variance explained, was .24, indicating that approximately 24% of the differences in scores among students were related to the classroom to which they were assigned.

Table 3  
*Descriptive Statistics for Amount (Minimum) of Student Instruction Variables and Cross-Level Associations for Amount and Classroom Learning Environment (CLE) Ratings (From 1 [Low Quality] to 6 [High Quality])*

Variable	N	M	SD	Minimum	Maximum
TCM-SG-CF	315	0.82	1.68	0.00	15.05
TCM-SG-MF	315	9.32	9.56	0.00	42.11
CM-SG-MF	315	4.84	6.13	0.00	35.84
CM-SG-CF	315	0.54	1.31	0.00	6.86
CM-WC-CF	315	0.81	1.91	0.00	17.78
CM-WC-MF	315	6.31	6.98	0.00	29.30
TCM-WC-CF	315	2.80	5.18	0.00	31.05
TCM-WC-MF	315	12.61	16.15	0.00	67.81

Note. Hierarchical linear modeling (HLM) cross-level student and classroom associations unstandardized coefficients.

Student-level outcome	Classroom-level CLE quality HLM coefficient (SE)
TCM-SG-MF	.261 (1.67)
TCM-WC-MF	-.386 (1.20)
TCM-SM-CF	.060 (.04)
TCM-WC-CF	-.035 (.23)
CM-SM-MF	.117 (.44)

Note. TCM = teacher/child-managed; SG = small group; CF = code-focused instruction; MF = meaning-focused instruction; CM = child-managed; WC = whole class.

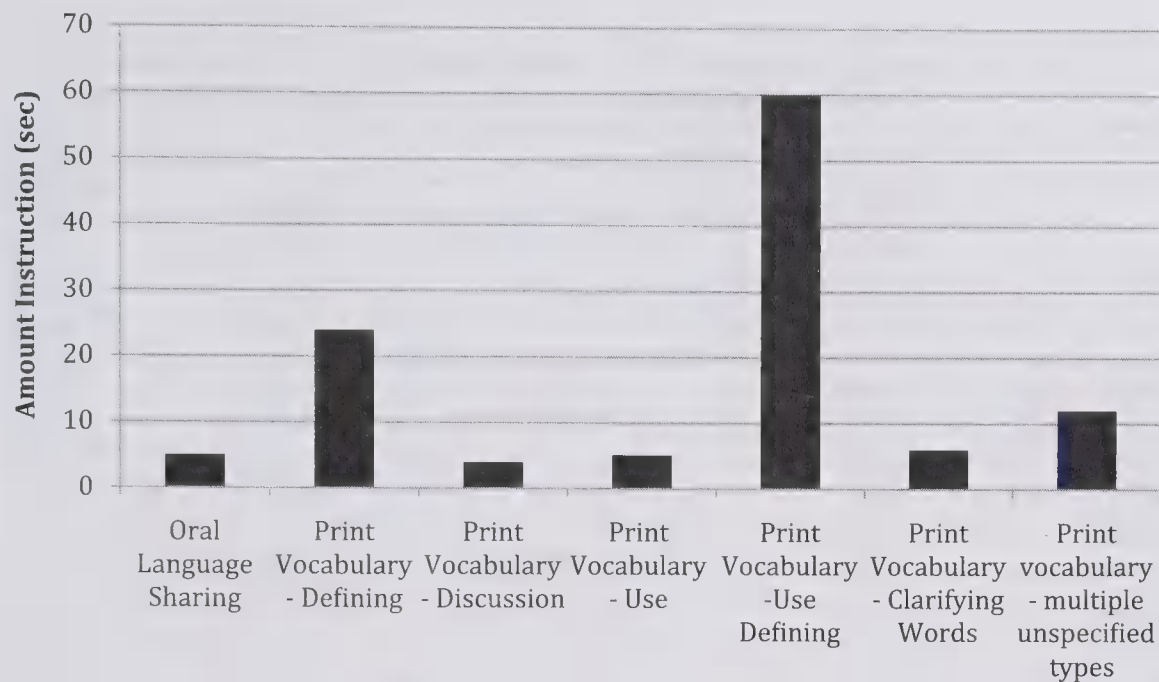


Figure 1. Amounts (s/day) of teacher/child-managed (TCM) small group vocabulary instruction by type.

### Main Effects of CLE and Amounts, Types, and Content of Instruction on Students' Vocabulary and Reading Comprehension Outcomes

We then examined the effect of the CLE score on students' spring VocComp score. There was a trend that CLE predicted students' spring outcomes ( $p = .064$ ). A 1-point increase in CLE quality score was associated with about a .02 increase in the VocComp spring score ( $d = .044$ , which is negligible). The model explained about 81% of the variability in children's scores compared with the unconditional model.

Next, we removed the CLE quality score from the model and added all eight of the student amount/type/content of instruction time. Results revealed that none of the instruction duration variables predicted students' scores. The model explained approximately 80% of the variance in student scores compared with the unconditional model.

### CLE Quality $\times$ Duration/Type/Content Interactions

We then added the CLE quality score to the model and tested all of the CLE Quality  $\times$  Time/Type/Content of Instruction interactions (see Table 4). We trimmed three variables (i.e., teacher/child-managed code-focused small group and whole-class instruction, and child-managed small group code-focused instruction) and the interactions that did not significantly predict student outcomes. We also tested for Student Fall VocComp  $\times$  Instruction interaction effects. None significantly predicted spring VocComp outcomes (e.g., Fall VocComp  $\times$  CLE Quality interaction effect coefficient =  $-.008$ ,  $p = .400$ ) and so were trimmed from the model. This model explained about 81.3% of the variability in students' scores.

Our models revealed a number of global CLE Quality  $\times$  Individual Student-Level Instruction Amount/Content interactions that significantly predicted students' outcomes (see Table 4 and Figure

2) regardless of students' fall VocComp  $z$  score. Specifically, for teacher/child-managed small group *and* whole-class meaning-focused instruction, when provided by teachers who were judged to be providing a higher quality CLE, the effect for students who spent more time in meaning-focused instruction was much greater, minute for minute, than the same amount of instruction provided by teachers who were judged to be providing a CLE of lesser quality. The effect was substantial. For example, a student who received 18 min of teacher/child-managed small group meaning-focused instruction and whose teachers received a CLE score of 17 (75th percentile of the sample) would achieve scores that were .43  $z$ -score points ( $M = 0$ ,  $SD = 1$ ) higher than a student who received the same amount of instruction but whose teachers received a rating of 13 (25th percentile of the sample), with an effect size ( $d$ ) of .43. The difference for teacher/child-managed whole-class meaning-focused instruction was smaller, .19  $z$ -score points ( $d = .19$ ), or about half of the small group effect size.

### Discussion

Overall, the teachers in this study provided generally high-quality CLEs, but there was substantial variability with two teachers providing CLEs that were judged to be very low—a 6 or worse (out of 18); they received no more than a 1, 2, or 3 on all three scales. At the same time, six teachers had almost perfect scores of 15, having received 5s and 6s on all three scales. There was similar variability in the amount of time third graders spent in teacher/child-managed meaning-focused instruction both within and between classrooms. Neither CLE quality nor the amount/content/type of instruction (ISI/Pathways) individual students received independently predicted students' vocabulary and comprehension gains. Instead, the two systems synergistically captured the complexity of classroom instruction at the individual student level and the more global classroom level. Teachers judged to provide a high-quality CLE but whose students received very little teacher/



Table 4

*HLM Results: Effect of Classroom Learning Environment (CLE), Amounts/Types/Content of Student Instruction (Minimum), and Interaction Effects on Students' Spring Vocabulary and Comprehension (VocComp) Z Scores and HLM Model*

Fixed effect	Coefficient	SE	<i>t</i>	Approx. <i>df</i>	<i>p</i>
Mean VocComp, $\gamma_{00}$	-0.034	0.032	-1.056	25	.301
CLE, $\gamma_{01}$	0.027	0.008	3.254	25	.003
For TCM-SG MF slope, $\beta_1$					
TCM-SG-MF effect, $\gamma_{10}$	-0.003	0.003	-0.751	331	.453
CLE interaction effect, $\gamma_{11}$	0.005	0.001	3.509	331	<.001
For CM-SG-MF slope, $\beta_2$					
CM-SG-MF effect, $\gamma_{20}$	0.005	0.005	0.971	331	.332
CLE interaction effect, $\gamma_{21}$	-0.003	0.001	-1.873	331	.062
For CM-WC-CF slope, $\beta_3$					
CM-WC-CF effect, $\gamma_{30}$	0.013	0.012	1.102	331	.271
CLE interaction effect, $\gamma_{31}$	-0.004	0.002	-1.897	331	.059
For CM-WC-MF slope, $\beta_4$					
CM-WC-MF effect, $\gamma_{40}$	0.001	0.004	0.314	331	.754
CLE interaction effect, $\gamma_{41}$	0.002	0.001	1.809	331	.071
For TCM-WC-MF slope, $\beta_5$					
TCM-WC-MF effect, $\gamma_{50}$	0.0001	0.001	0.156	331	.876
CLE interaction effect, $\gamma_{51}$	0.001	0.000	2.816	331	.005
For fall VocComp slope, $\beta_6$					
Fall VocComp effect, $\gamma_{60}$	0.901	0.027	32.755	331	<.001
Random effect	SD	Variance component	<i>df</i>	$\chi^2$	<i>p</i>
Classroom level, $u_0$	0.14574	0.02124	25	50.16340	.002
Student level, $r$	0.41352	0.17100			
Deviance = 464.152946					

*Note.* HLM = hierarchical linear modeling; TCM = teacher/child-managed; SG = small group; MF = meaning-focused; CM = child-managed; WC = whole class; CF = code-focused instruction. Model:  $Spring\ CompVoc_{ij} = \gamma_{00} + \gamma_{01} * CLE_j + \gamma_{10} * TCM_{SGMF}_{ij} + \gamma_{11} * CLE_j * TCM_{SGMF}_{ij} + \gamma_{20} * CM_{SGMF}_{ij} + \gamma_{21} * CLE_j * CM_{SGMF}_{ij} + \gamma_{30} * CM_{WC-CF}_{ij} + \gamma_{31} * CLE_j * CM_{WC-CF}_{ij} + \gamma_{40} * CM_{WC-MF}_{ij} + \gamma_{41} * CLE_j * CM_{WC-MF}_{ij} + \gamma_{50} * TCM_{WC-MF}_{ij} + \gamma_{51} * CLE_j * TCM_{WC-MF}_{ij} + \gamma_{60} * Fall\ CompVoc\ Factor\ Score_{ij} + \gamma_{61} * CLE_j * Fall\ CompVoc_{ij} + u_{0j} + r_{ij}$ .

child-managed meaning-focused instruction (e.g., less than 1 min, see Figure 2) were no more effective than those judged to provide a low-quality CLE. As students spent more time in teacher/child-managed meaning-focused small group and whole-class instruction, differences in low- versus high-quality CLE effects became larger. Moreover, because students who shared a classroom experienced different amounts of small group teacher/child-managed meaning-focused instruction, using both systems helped to explain within-classroom differences in students' outcomes, which classroom-level systems obscure.

Students showed the greatest gains in vocabulary and comprehension when their teachers provided a high-quality CLE and they spent greater amounts of time (e.g., 25–35 min, see Figure 2) in teacher/child-managed meaning-focused instruction, particularly when this instruction was provided in small groups. Teacher/child-managed small group meaning-focused instruction was more than twice as effective as whole-class instruction. When we tested students' Fall Score  $\times$  Instruction interactions, none significantly predicted spring outcomes. This indicates that results were similar for students regardless of fall vocabulary and comprehension scores.

Another consideration is that for this sample of students, teachers, and classrooms, Student Characteristic  $\times$  Instruction Time/Type/Content effects on students' literacy gains were documented (Connor, Morrison, Fishman, et al., 2011). Specifically, using the ISI/Pathways system, the smaller the difference between the observed amount of a particular type/content of instruction and the recommended amount for a particular student based on his or her

assessed vocabulary and literacy skills (i.e., distance from recommendation), the greater were his or her reading comprehension skill gains.

These findings support a complex systems model of how individual children learn in the context of classroom literacy instruction and, in combination with other studies (Connor, Morrison, Fishman, et al., 2011), extend our understanding of classroom instruction and learning environments as dynamic systems with interacting effects (Yoshikawa & Hsueh, 2001). This indicates that we will be more likely to identify key aspects of complex teaching and CLEs by using multiple frameworks and considering potential interactions among sources of influence at both the more micro student level as well as the more global classroom level. Thus, a student in a CLE that is generally judged to be of high quality may not show achievement gains because the student is not receiving appropriate amounts of the particular types and content of instruction that would support his or her achievement. Indeed, students in this study who were judged to be in a high-quality CLE were no more likely to participate in substantial amounts of teacher/child-managed meaning-focused instruction than were students in a low-quality CLE. As one reviewer noted, "These are the well-organized, 'nice' classrooms that [some] principals and parents love, even though the environment while pleasant is not particularly effective."

Such interaction effects may help to explain equivocal findings when only one system of observation is used and individual student differences are not considered. For example, in the Kane et al. (2012) study, one reason that the assessments of CLE might not

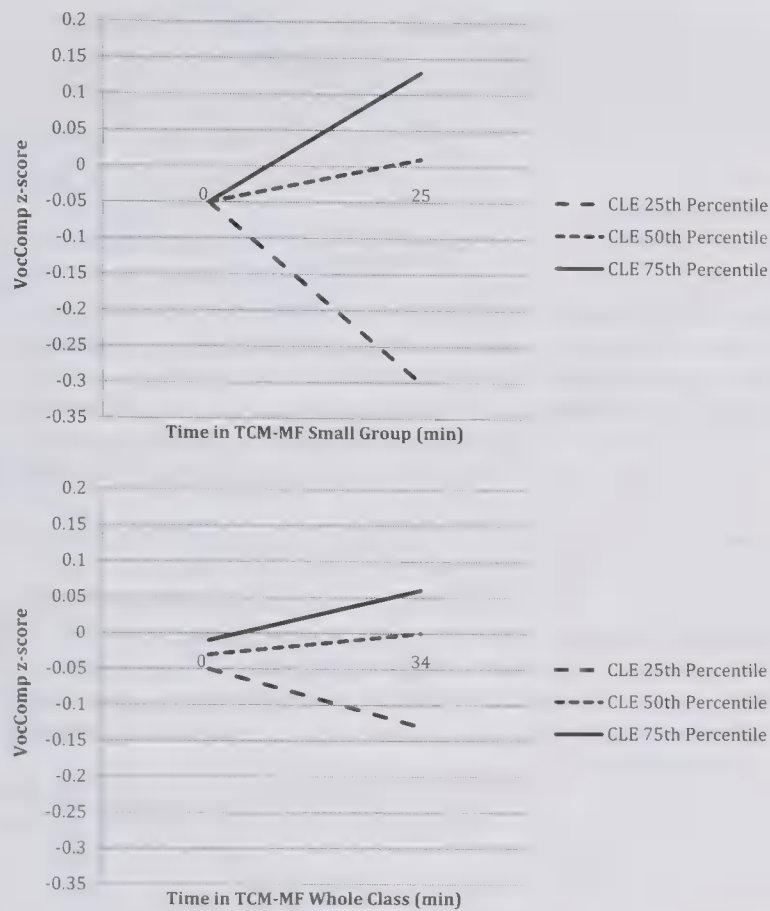


Figure 2. Student-level Amounts, Content, and Type of Instruction Received  $\times$  Quality of the Classroom Learning Environment (CLE) interactions on Vocabulary/Comprehension (VocComp) z scores. Modeled x-axis 5th to 95th percentile of time spent in teacher/child-managed-meaning-focused (TCM-MF) small group (top) and whole-class (bottom) instruction, as a function of CLE score modeled at the 25th, 50th, and 75th percentiles; other variables are centered at their mean.

have predicted language arts outcomes was because, even in classrooms with teachers judged to be of high quality, students might not have received adequate amounts of specific types of content instruction tailored to their learning needs. None of the observation systems used in the Kane study considered time in content and type of literacy instruction at the student level or Student Characteristic  $\times$  Instruction interaction effects. At the same time, systems that consider a single time point in certain types-content of literacy instruction and that do not consider quality or Child  $\times$  Instruction interactions may not be generally highly predictive either because more time spent in low-quality unaligned instruction is unlikely to positively predict student achievement.

On average, students spent about only 5 min per day engaged in oral language and vocabulary instruction, which, by any standard, is a minimal amount of time. This implies that there was very little time for explicit instruction or other types of vocabulary instruction (e.g., discussing or clarifying words) to take place. These findings are in line with previous research, which has demonstrated that in many elementary school classrooms, there is a very limited focus on vocabulary instruction (Biemiller, 2001; Durkin, 1979; Scott & Nagy, 1997). Robust vocabulary instruction (Beck et al., 2002) requires teachers to allot sufficient amounts of time that can be used for explaining, providing examples, and elaborating on vocabulary knowledge in order to promote greater understanding.

Five minutes per day of vocabulary instruction seems hardly adequate.

A closer look at the types of comprehension instruction provided (see Appendix C) revealed that teachers used over 20 different kinds of teacher/child-managed whole-class comprehension instructional activities (see Table C.1) and fewer (about 18) kinds of activities in small groups (Table C.2) to build comprehension. The most salient types observed involving the whole class were *questioning* (3.1 min), *highlighting/identifying* (about 1 min), and *schema building* (about 45 s) (see Appendix B). In teacher/child-managed small group instruction, again, the most salient activity observed was *questioning* (1.6 min), followed by *compare and contrast* (13 s) and the use of *graphic organizers* (8 s). In both whole-class and small group contexts, amounts varied widely (see Table C.1). For example, *questioning* ranged from 0 to 22 min, and use of *graphic organizers* ranged from 0 to 16 min.

In general, the comprehension instruction observed was aligned with the findings of the National Reading Panel report (NICHD, 2000), with a focus on strategies (compare and contrast; highlighting; graphic organizers) but little support for more complex understanding of text, which is now required by the Common Core State Standards. We consider questioning among the most basic tools the teacher might use to build reading for understanding (Cazden, 1988). In contrast, very little time was spent on higher level inferencing in either whole class (about 15 s summing across types for whole class) or small group (about 12 s). Research shows that inferencing is associated with successful comprehension (Cain et al., 2004; Cromley & Azevedo, 2007) and is a core principal of the Common Core State Standards (Common Core State Standards Initiative, 2010).

One aspect of the study that deserves further investigation is whether dimensions of the CLE might be better predictors when coded at the level of the individual student. It is conceivable that Student A and Student C might be participating in the same amount of time in appropriate learning opportunities but that the teacher is interacting with Student C in ways that are more responsive than with Student A. Hence, one might hypothesize that Student C will demonstrate stronger achievement than will Student A. Measuring CLE and instruction at the level of the individual student is time-consuming but may be worth the effort when trying to understand learning and development in classrooms. Another consideration is that dimensions of the CLE quality measure captured the social and emotional climate of the classroom as well as the learning environment. It may be that, particularly for children from low-income families, this more nurturing aspect of the classroom environment is providing a safe haven that facilitates learning, albeit indirectly. There is evidence of this effect in preschool, kindergarten, and first-grade classrooms (Hamre & Pianta, 2005; Rimm-Kaufman et al., 2005) as well as for older children (Reyes, Brackett, Rivers, White, & Salovey, 2012).

## Limitations

There are limitations to this study that should be considered when interpreting these results. First, all participating teachers received professional development. This study was conducted in the context of a randomized controlled trial in which there was a significant effect of the individualized reading compared with the



vocabulary intervention (Connor, Morrison, Fishman, et al., 2011). Although there were no differences between conditions on the quality of the CLE, students who were in classrooms where teachers learned to individualize instruction were more likely to participate in teacher/child-managed small group meaning-focused instruction that was individualized to their learning needs (Connor, Morrison, Fishman, et al., 2011). The randomized control trial may have influenced the results presented here, and it is possible that our results may not generalize to classrooms in which teachers do not receive professional development. It might also explain why there were no significant Student Fall VocComp Score  $\times$  Instruction interaction effects on spring outcomes. In another sample, such interaction effects might influence achievement. Additionally, the three observations were scheduled at the teachers' convenience, so we most likely observed higher quality instruction than might have been observed otherwise. Use of more frequent observations would have improved reliability (Rowan & Correnti, 2009) but were not feasible within the funding and time constraints of this study.

### Practical Implications

This study provides insight into the amounts, content, and types of instruction in which individual students participate and qualitative aspects of the CLE that appear to be more effective for improving students' literacy and language outcomes. Teacher/child-managed meaning-focused but not teacher/child-managed code-focused instruction predicted third graders' vocabulary and comprehension achievement. This might not be unexpected inasmuch as explicit instruction in the target outcome tends to be a better predictor than more implicit or indirect instruction (Connor, Morrison, & Katch, 2004; Connor, Morrison, & Petrella, 2004; Connor, Morrison, & Slominski, 2006), and the outcomes were specifically meaning-focused. Small group instruction was twice as effective as whole class, perhaps because the teacher was better able to tailor instruction to meet the learning needs of students when interacting with smaller numbers of students. Additionally, he or she was likely to be more responsive, which the extant literature has established as important for student learning (Mashburn et al., 2008). This responsiveness was a key dimension of the CLE rubric. Perhaps, the most important finding was that type, amount, and content of instruction individual students received and the quality of the CLE matter: Students should learn best when provided enough time in explicit instruction from the teacher who is interactive, responsive, organized, and focused on providing targeted language and literacy content in ways that facilitate language and vocabulary learning.

Classroom observation systems represent an important move toward policy that promises to make a true difference in what is defined as high-quality and effective teaching, what it looks like in the classroom, and how these practices can be more widely disseminated so that all students, including the most vulnerable, can experience effective instruction and academic gains. One challenge will be designing systems that can be used validly and reliably by school professionals (Crawford et al., in press; Reddy et al., in press) who have varying levels of expertise and knowledge about literacy development. By better understanding the affordances of teaching and the CLE that contribute to individual student's language and literacy development, we can design more

effective instructional regimens, identify effective standards of practice, discover better ways to measure effective teaching, and develop targeted professional development for teachers and educational leaders that will ensure that all children have the opportunity to learn.

### References

- Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction. Solving problems in the teaching of literacy*. New York, NY: Guilford Press.
- Beck, I. L., Perfetti, C. A., & McKeown, M. G. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology*, 74, 506–521. doi:10.1037/0022-0663.74.4.506
- Biemiller, A. (2001). Teaching vocabulary: Early, direct, and sequential. *American Educator*, 25(1), 24–28.
- Biemiller, A., & Boote, C. (2006). An effective method for building meaning vocabulary in primary grades. *Journal of Educational Psychology*, 98, 44–62. doi:10.1037/0022-0663.98.1.44
- Bogdan, R. C., & Biklen, S. K. (1998). *Qualitative research in education: An introduction to theory and method* (3rd ed.). Needham Heights, MA: Allyn & Bacon.
- Borko, H., & Niles, J. (1987). Descriptions of teacher planning: Ideas for teachers and research. In V. Richardson-Koehler (Ed.), *Educators' handbook: A research perspective* (pp. 167–187). New York, NY: Longman.
- Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In R. M. Lerner & W. Damon (Eds.), *Handbook of child psychology: Theoretical models of human development* (6th ed., Vol. 1, pp. 793–828). Hoboken, NJ: John Wiley & Sons.
- Brophy, J. E. (1979). Teacher behavior and its effects. *Journal of Educational Psychology*, 71, 733–750. doi:10.1037/0022-0663.71.6.733
- Cain, K., Oakhill, J., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology*, 96, 671–681. doi:10.1037/0022-0663.96.4.671
- Cameron, C. E., Connor, C. M., & Morrison, F. J. (2005). Effects of variation in teacher organization on classroom functioning. *Journal of School Psychology*, 43, 61–85. doi:10.1016/j.jsp.2004.12.002
- Cameron, C. E., Connor, C. M., Morrison, F. J., & Jewkes, A. M. (2008). Effects of classroom organization on letter–word reading in first grade. *Journal of School Psychology*, 46, 173–192. doi:10.1016/j.jsp.2007.03.002
- Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading & Writing*, 12, 169–190. doi:10.1023/A:1008131926604
- Cassidy, J., & Cassidy, D. (2005/2006). What's hot, what's not for 2006. *Reading Today*, 23, 1.
- Cazden, C. (1988). *Classroom discourse*. Portsmouth, NH: Heineman.
- Chall, J. S. (1967). *Learning to read: The great debate*. New York, NY: McGraw-Hill.
- Common Core State Standards Initiative. (2010). *Common Core State Standards for Mathematics*. Retrieved from [http://www.corestandards.org/assets/CCSSI\\_MathStandards.pdf](http://www.corestandards.org/assets/CCSSI_MathStandards.pdf)
- Connor, C. M. (2013a). Commentary on two classroom observation systems: Moving toward a shared understanding of effective teaching. *School Psychology Quarterly*, 28, 342–346. doi:10.1037/spq0000045
- Connor, C. M. (2013b). Intervening to support reading comprehension development with diverse learners. In B. Miller & L. E. Cutting (Eds.), *Unraveling the behavioral, neurobiological and genetic components of reading comprehension: The Dyslexia Foundation and NICHD* (pp. 222–232). Baltimore, MD: Brookes.

- Connor, C. M., & Morrison, F. J. (2012). Knowledge acquisition in the classroom: Literacy and content area knowledge. In A. M. Pinkham, T. Kaefer, & S. B. Neuman (Eds.), *Knowledge development in early childhood: How young children build knowledge and why it matters* (pp. 220–241). New York, NY: Guilford Press.
- Connor, C. M., Morrison, F. J., Fishman, B., Crowe, E. C., Al Otaiba, S., & Schatschneider, C. (2013). A longitudinal cluster-randomized control study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. *Psychological Science*, 24, 1408–1419. doi:10.1177/0956797612472204
- Connor, C. M., Morrison, F. J., Fishman, B., Giuliani, S., Luck, M., Underwood, P., . . . Schatschneider, C. (2011). Classroom instruction, child X instruction interactions and the impact of differentiating student instruction on third graders' reading comprehension. *Reading Research Quarterly*, 46, 189–221.
- Connor, C. M., Morrison, F. J., Fishman, B., Ponitz, C. C., Glasney, S., Underwood, P., . . . Schatschneider, C. (2009). The ISI classroom observation system: Examining the literacy instruction provided to individual students. *Educational Researcher*, 38, 85–99.
- Connor, C. M., Morrison, F. J., & Katch, L. E. (2004). Beyond the reading wars: Exploring the effect of child-instruction interactions on growth in early reading. *Scientific Studies of Reading*, 8, 305–336.
- Connor, C. M., Morrison, F. J., & Petrella, J. N. (2004). Effective reading comprehension instruction: Examining Child X Instruction interactions. *Journal of Educational Psychology*, 96, 682–698.
- Connor, C. M., Morrison, F. J., Schatschneider, C., Toste, J., Lundblom, E. G., Crowe, E., & Fishman, B. (2011). Effective classroom instruction: Implications of child characteristic by instruction interactions on first graders' word reading achievement. *Journal for Research on Educational Effectiveness*, 4, 173–207.
- Connor, C. M., Morrison, F. J., & Slominski, L. (2006). Preschool instruction and children's literacy skill growth. *Journal of Educational Psychology*, 98, 665–689.
- Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., . . . Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of child by instruction interactions on first graders' literacy development. *Child Development*, 80, 77–100.
- Connor, C. M., Son, S.-H., Hindman, A. H., & Morrison, F. J. (2005). Teacher qualifications, classroom practices, family characteristics, and preschool experience: Complex effects on first graders' vocabulary and early reading outcomes. *Journal of School Psychology*, 43, 343–375.
- Crawford, A. D., Zucker, T. A., Williams, J. M., Bhavsar, V., & Landry, S. H. (in press). Initial validation of the pre-kindergarten classroom observation tool and goal setting system for data-based coaching. *School Psychology Quarterly*.
- Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, 99, 311–325. doi:10.1037/0022-0663.99.2.311
- Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.
- Durkin, D. (1979). *Teaching them to read*. Boston, MA: Allen & Bacon.
- Goldhaber, D., & Anthony, E. (2003). *Teacher quality and student achievement. Urban Diversity Series* (Report: UDS-115; 153). New York, NY: Department of Education, Washington, DC.
- Gottardo, A., Stanovich, K. E., & Siegel, L. S. (1996). The relationships between phonological sensitivity, syntactic processing, and verbal working memory in the reading performance of third-grade children. *Journal of Experimental Child Psychology*, 63, 563–582. doi:10.1006/jecp.1996.0062
- Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2010). *Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores* (No. 45). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research. doi:10.3386/w16015
- Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development*, 76, 949–967. doi:10.1111/j.1467-8624.2005.00889.x
- Hamre, B. K., Pianta, R. C., Downer, J. T., & Mashburn, A. J. (2007, April). *Growth models of classroom quality over the course of the year in preschool programs*. Paper presented at the biennial meeting of the Society for Research in Child Development, Boston, MA.
- Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, 26, 430–511. doi:10.1080/07370000802177235
- Justice, L. M., Petscher, Y., Schatschneider, C., & Mashburn, A. (2011). Peer effects in preschool classrooms: Is children's language growth associated with their classmates' skills. *Child Development*, 82, 1768–1777. doi:10.1111/j.1467-8624.2011.01665.x
- Kane, T., Staiger, D. O., & McCaffrey, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Retrieved from <http://www.metproject.org>
- Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, 12, 281–300. doi:10.1080/10888430802132279
- Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York, NY: Guilford Press.
- Konstantopoulos, S., & Chung, N. (2011). The persistence of teacher effects in elementary grades. *American Educational Research Journal*, 48, 361–386. doi:10.3102/0002831210382888
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174. doi:10.2307/2529310
- MacGinitie, W. H., & MacGinitie, R. K. (2006). *Gates-MacGinitie Reading Tests* (4th ed.). Iowa City, IA: Houghton Mifflin.
- Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., . . . Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development*, 79, 732–749. doi:10.1111/j.1467-8624.2008.01154.x
- Mehta, P. D., Foorman, B. R., Branum-Martin, L., & Taylor, W. P. (2005). Literacy as a unidimensional multilevel construct: Validation, sources of influence, and implications in a longitudinal study in grades 1 to 4. *Scientific Studies of Reading*, 9, 85–116. doi:10.1207/s1532799xssr0902\_1
- National Center for Education Statistics, Institute of Education Sciences, US Department of Education. (2013). *Fast facts: English language learners*. Retrieved from Institute of Education Sciences, US Department of Education <http://nces.ed.gov/fastfacts/display.asp?id=96>
- National Institute of Child Health and Human Development. (2000). *National Institute of Child Health and Human Development, National Reading Panel report: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, Author.
- Noldus, L. P. J. J., Trienes, R. J. H., Hendriksen, A. H. M., Jansen, H., & Jansen, R. G. (2000). The Observer Video-Pro: New software for the collection, management, and presentation of time-structured data from videotapes and digital media files. *Behavior Research Methods, Instruments, & Computers*, 32, 197–206.
- Pianta, R. C., Belsky, J., Houts, R., & Morrison, F. (2007). Opportunities to learn in America's elementary classrooms. *Science*, 315, 1795–1796.



- Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS) manual, K-3*. Baltimore, MD: Brookes.
- Pianta, R. C., La Paro, K. M., Payne, K., Cox, C., & Bradley, R. H. (2002). The relation of kindergarten classroom environment to teacher, family and school characteristics and child outcomes. *Elementary School Journal, 102*, 225–238. doi:10.1086/499701
- Piasta, S. B., Connor, C. M., Fishman, B., & Morrison, F. J. (2009). Teachers' knowledge of literacy, classroom practices, and student reading growth. *Scientific Studies of Reading, 13*, 224–248. doi:10.1080/10888430902851364
- Ramey, S. L., & Ramey, C. (2006). Early educational interventions: Principles of effective and sustained benefits from targeted early education programs. In D. K. Dickinson & S. B. Neuman (Eds.), *Handbook of early literacy research* (Vol. 2, pp. 445–459). New York, NY: Guilford Press.
- Rasch, G. (2001). Winsteps (Version 3.30) [Statistics]. Retrieved from <http://www.winsteps.com/winsteps.htm>
- Raudenbush, S. W. (2009). The Brown legacy and the O'Connor challenge: Transforming schools in the images of children's potential. *Educational Researcher, 38*, 169–180. doi:10.3102/0013189X09334840
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Raver, C. C., Blair, C., & Li-Grining, C. P. (2011). *Extending models of emotion self-regulation to classroom settings: Implications for professional development*. Manuscript submitted for publication.
- Reddy, L. A., Fabiano, G., Dudek, C., & Hsu, L. (in press). Predictive validity of the Classroom Strategies Scale-Observer Form on statewide testing scores: An initial investigation. *School Psychology Quarterly*.
- Reis, S. M., McCoach, D. B., Little, C. A., Muller, L. M., & Kaniskan, R. B. (2011). The effects of differentiated instruction and enrichment pedagogy on reading achievement in five elementary schools. *American Educational Research Journal, 48*, 462–501. doi:10.3102/0002831210382891
- Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology, 104*, 700–712. doi:10.1037/a0027268
- Reynolds, A. J., Magnuson, K. A., & Ou, S.-R. (2010). Preschool-to-third grade programs and practices: A review of research. *Children and Youth Services Review, 32*, 1121–1131. doi:10.1016/j.chilcyouth.2009.10.017
- Rimm-Kaufman, S. E., La Paro, K. M., Downer, J. T., & Pianta, R. C. (2005). The contribution of classroom setting and quality of instruction to children's behavior in kindergarten classrooms. *Elementary School Journal, 105*, 377–394.
- Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the study of instructional improvement. *Educational Researcher, 38*, 120–131. doi:10.3102/0013189X09332375
- Rupley, W. H., Logan, J. W., & Nichols, W. D. (1998). Vocabulary instruction in a balanced reading program. *Reading Teacher, 52*, 336–346.
- Scott, J. A., & Nagy, W. E. (1997). Understanding the definitions of unfamiliar verbs. *Reading Research Quarterly, 32*, 184–200.
- Snow, C. E. (2001). *Reading for understanding*. Santa Monica, CA: RAND Education and the Science and Technology Policy Institute.
- Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.
- Stahl, S. A. (1999). *Vocabulary development*. Cambridge, MA: Brookline Books.
- Taylor, B. M., & Pearson, D. P. (Eds.). (2002). *Teaching reading: Effective schools, accomplished teachers*. Mahwah, NJ: Lawrence Erlbaum.
- Taylor, B. M., Pearson, D. P., Clark, K., & Walpole, S. (2000). Effective schools and accomplished teachers: Lessons about primary-grade reading instruction in low-income schools. *Elementary School Journal, 101*, 121–165. doi:10.1086/499662
- Therrien, W. J. (2004). Fluency and comprehension gains as a result of repeated reading: A meta-analysis. *Remedial and Special Education, 25*, 252–261. doi:10.1177/07419325040250040801
- Tuyay, S., Jennings, L., & Dixon, C. (1995). Classroom discourse and opportunities to learn: An ethnographic study of knowledge construction in a bilingual third-grade classroom. *Discourse Processes, 19*, 75–110. doi:10.1080/01638539109544906
- UTeach Teacher Observer Protocol. (2009). *Development of the UTeach Observation Protocol: A classroom observation instrument to evaluate mathematics and science teachers from the UTeach Preparation Program*. Retrieved October 1, 2012, from [https://uteach.utexas.edu/sites/default/files/UTOP\\_Paper\\_Non\\_Anonymous\\_4\\_3\\_2011.pdf](https://uteach.utexas.edu/sites/default/files/UTOP_Paper_Non_Anonymous_4_3_2011.pdf)
- Vellutino, F. R., Tunmer, W. E., Jaccard, J., Chen, R., & Scanlon, D. M. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific Studies of Reading, 11*, 3–32. doi:10.1207/s1532799xssr1101\_2
- Wharton-McDonald, R., Pressley, M., & Hampston, J. M. (1998). Literacy instruction in nine first-grade classrooms: Teacher characteristics and student achievement. *Elementary School Journal, 99*, 101–128. doi:10.1086/461918
- Whitehurst, G. J., Falco, F. L., Lonigan, C. J., Fischel, J. E., DeBaryshe, B. D., Valdez-Menchaca, M. C., & Caulfield, M. (1988). Accelerating language development through picture book reading. *Developmental Psychology, 24*, 552–559. doi:10.1037/0012-1649.24.4.552
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson-III Tests of Achievement*. Itasca, IL: Riverside.
- Yoshikawa, H., & Hsueh, J. (2001). Child development and public policy: Toward a dynamic systems perspective. *Child Development, 72*, 1887–1903. doi:10.1111/1467-8624.00384

## Appendix A

### The ISI Classroom Learning Environment Scale (Excerpts)

Classroom Orienting, Organization and Planning	Support for Language/Vocabulary	Warmth and Responsiveness/Control/Discipline
<p><b>Rating 1 Indicators</b> No evidence of classroom organization. Teacher frequently does not have materials ready or enough materials for all children. Classroom is frequently chaotic and very little time is spent on meaningful instruction. No observable system is in place to facilitate students' transition from one station or location to another.</p>	<p><b>Rating 1 Indicators</b> Teacher does not introduce any new words, does not provide explicit or systematic instruction in vocabulary and does not provide opportunities for students to engage in oral language. Students are not provided opportunities to practice key vocabulary. Teacher does not monitor students' vocabulary and comprehension.</p>	<p><b>Rating 1 Indicators</b> No evidence that the teacher redirects in respectful ways, nor is there evidence that the teacher emphasizes student change in behavior through praise. There is no evidence of the teacher communicating what students did correctly or how they can improve. There is no evidence of students treating each other with respect. Whenever discipline is imposed, it is ineffective.</p>
<p><b>Rating 3 Indicators</b> Transitions are of reasonable length but not consistently efficient (not all children). There is an observable, but not always efficient or working system (e.g., center chart, daily schedule) in place for organizing students into groups. The teacher may use a daily lesson plan (e.g., group activity planner print-out).</p>	<p><b>Rating 3 Indicators</b> Teacher introduces too many new Tier 1 words per story/text and not enough Tier 2 words. Provides explicit or systematic vocabulary instruction (not both). Occasionally extends meanings. Occasionally provides opportunities for students to engage in oral language and practice key vocabulary. Teacher monitors students' vocabulary and comprehension, but rarely provides feedback.</p>	<p><b>Rating 3 Indicators</b> Teacher inconsistently redirects in respectful ways and inconsistently emphasizes student change in behavior through praise. Teacher talk is inconsistently encouraging and respectful and inconsistently connects students' personal experiences to lesson content. Inconsistently communicates clearly what students did correctly or how they can improve. Students inconsistently treat each other with respect.</p>
<p><b>Rating 6 Indicators</b> The classroom is well organized and instruction is well organized. Classroom routine is evident. Transitions are efficient.</p>	<p><b>Rating 6 Indicators</b> Word knowledge is an ongoing part of the instructional day. Teacher's selection of words demonstrates knowledge of the words' utility and relation to previously known words and relevance for text being taught. Students are encouraged to make connections between words/meaning they are already familiar with and new words/meanings. Much of the instruction is provided in small groups.</p>	<p><b>Rating 6 Indicators</b> Teacher is the authority figure in the class but is never punitive. Classroom consistently offers a positive learning environment with clear expectations for students' behavior as a member of the learning community. Effectively selects and incorporates students' responses, ideas, examples, and experiences into the lesson.</p>

*Note.* The full scale is available upon request. ISI = Individualizing Student Instruction.

(Appendices continue)



## Appendix B

### Listening and Reading Comprehension Excerpts From the ISI/Pathways Coding Manual (full manual available upon request)

Comprehension should be coded for activities intended to increase students' comprehension of written or oral text. This includes instruction and practice in using comprehension strategies and demonstration of comprehension abilities. Comprehension activities generally follow or are incorporated into reading or listening of connected text (e.g., silent sustained reading followed by a comprehension worksheet, comprehension strategy instruction using a particular example of connected text, an interactive teacher read aloud during which the teacher models various comprehension strategies).

**7.1.10.3 Schema and Concept Building (Modifier).** Listening and Reading Comprehension>Schema Building should be coded for activities which involve the **teacher** clarifying a concept and building background knowledge. For example, the teacher tells the students about the middle ages while reading a fairy tale. Discussions about specific words should be coded as Print Vocabulary>Class Discussion.

**7.1.10.4 Predicting (Modifier).** Predicting should be coded for activities which involve predicting future events or information not yet presented based on information already conveyed by the text (e.g., making predictions from foreshadowing). Predicting occurs while reading a story and involves specific details or events, as opposed to Comprehension>Previewing, which involves a general prediction of what the text will be about.

**7.1.10.6 Inferencing – Within-Texts (Modifier).** Inferencing-Within-Texts should be coded for activities that involve making inferences within a text based on information that has not been explicitly stated in the text, but is inferred from information already conveyed in the text. An example of this would be if the students were reading a story about a boy who lost his dog and the teacher asks the students, "How do you think the boy felt when he finally found his dog at the end of the story?"

**7.1.10.7 Inferencing – Background Knowledge (Modifier).** Inferencing – Background Knowledge should be coded for activ-

ities that involve making inferences within a text based on information that has not been explicitly stated in the text but is based on activating students' background knowledge to make connections between their own knowledge/experiences and information presented in the text to make inferences about the story. An example of this would be if the teacher is reading a story about a boy who loses his dog and the teacher says, "Have any of you ever lost a pet? How did it make you feel? How do you think the boy in the story feels?" \*\* The difference between Inferencing-Background Knowledge vs. Prior Knowledge is that the teacher must *explicitly* ask the students to make an inference by activating background knowledge.

**7.1.10.8 Questioning (Modifier).** Listening and Reading Comprehension>Questioning should be coded for activities which involve generating or answering questions regarding factual or contextual knowledge from the text (e.g., What did Ira miss when he went to the sleepover? What was the name of \_\_\_\_\_?), provided that these activities are not better coded by Comprehension>Prior Knowledge (e.g., when the teacher uses a question to scaffold children in activating personal knowledge related to the text: "When you go to an amusement park, what do you expect to see?"), Comprehension>Monitoring (e.g., when the teacher uses a question aimed at stimulating students' metacognitive assessment of whether they comprehended the text: "Did I understand what happened there?"), or Comprehension>Predicting (e.g., when the teacher asks students to predict what will happen next: "What do you think the lost boy will do now?"). Questioning should also be coded for AR tests which are typically completed on the computer; AR test should also be coded as event code > Assessment. This code should also be used as a default code for activities where it is not clear whether activity is highlighting, questioning, or summarizing.

(Appendices continue)

## Appendix C

Table C1

*Descriptive Statistics for Teacher/Child-Managed Comprehension Whole-Class Instruction in Seconds*

	Min	Max	<i>M</i>	<i>SD</i>
Type of comprehension instruction				
Previewing	.00	374.73	24.42	72.80
Schema building	.00	588.29	46.60	112.47
Questioning	.00	1772.7	186.0	294.66
Monitoring	.00	235.49	7.15	29.59
Highlighting/identifying	.00	985.10	55.91	126.84
Context cues	.00	327.01	8.04	37.45
Graphic/semantic organizers	.00	725.37	12.54	86.33
Prior knowledge	.00	369.65	19.24	52.31
Retell	.00	192.24	7.75	29.95
Sequencing	.00	168.71	1.03	12.81
Compare/contrast	.00	380.58	22.77	73.03
Comprehension other	.00	847.11	17.72	70.53
Multicomponent integrated	.00	808.00	19.91	100.80
Comprehension strategies				
Predicting	.00	109.56	9.93	24.28
Inferencing between texts	.00	39.39	.01	4.46
Inferencing background knowledge	.00	180.55	5.38	23.98
Inferencing within text	.00	320.21	8.92	34.85
Summarizing main idea	.00	129.28	3.16	15.95
Fact vs. opinion	.00	159.19	2.06	18.05
Cause and effect	.00	327.01	3.98	31.47
Narrative text	.00	539.77	8.71	62.59
Expository text	.00	118.34	1.15	11.64

(Appendices continue)



Table C2  
*Descriptive Statistics for Teacher/Child-Managed Comprehension Small Group Instruction in Seconds*

	Min	Max	<i>M</i>	<i>SD</i>
Type of comprehension instruction				
Previewing	.00	266.06	2.10	18.60
Schema building	.00	248.71	6.13	27.88
Questioning	.00	1338.96	95.26	217.63
Monitoring	.00	111.81	0.36	6.37
Highlighting identifying	.00	169.00	7.22	27.02
Context cues	.00	178.70	2.49	17.17
Graphic/semantic organizers	.00	989.42	8.03	81.57
Prior knowledge	.00	103.43	2.09	9.42
Retell	.00	33.63	0.10	1.91
Sequencing	.00	36.39	0.28	2.90
Compare/contrast	.00	794.63	13.27	89.27
Multicomponent integrated	.00	420.09	6.30	41.95
Comprehension strategies				
Predicting	.00	569.08	9.03	65.10
Inferencing background knowledge	.00	333.89	2.74	26.83
Summarizing/main idea	.00	196.00	2.47	18.84
Fact vs. opinion	.00	629.78	5.67	51.18
Cause and effect	.00	176.17	1.08	13.43
Narrative text	.00	14.77	0.04	0.84
Workbook worksheet	.00	1016.78	9.57	77.19
Narrative Text Language Arts	.00	919.15	48.41	142.05
Expository Text Language Arts	.00	415.08	5.82	46.34
Expository Text Science	.00	560.82	11.59	65.63
Workbook Worksheet Language	.00	989.11	49.18	158.23
Workbook Worksheet Social Studies	.00	174.10	1.13	14.00
Blackboard Language Arts	.00	54.20	0.17	3.08
Expository Text Social Studies	.00	255.99	0.94	14.71
Workbook Worksheet Science	.00	575.05	10.32	60.75
Journal	.00	137.78	0.44	7.85
Summarizing	.00	196.00	3.28	19.80
Inferencing within text	.00	294.22	9.48	35.59

Received December 9, 2012

Revision received January 6, 2014

Accepted January 8, 2014 ■

# Text Comprehension Mediates Morphological Awareness, Syntactic Processing, and Working Memory in Predicting Chinese Written Composition Performance

Connie Qun Guan

University of Science and Technology Beijing

Feifei Ye

University of Pittsburgh

Richard K. Wagner

Florida State University

Wanjin Meng

China National Institute of Education Sciences

Che Kan Leong

University of Saskatchewan

The goal of the present study was to test opposing views about 4 issues concerning predictors of individual differences in Chinese written composition: (a) whether morphological awareness, syntactic processing, and working memory represent distinct and measureable constructs in Chinese or are just manifestations of general language ability; (b) whether they are important predictors of Chinese written composition and, if so, the relative magnitudes and independence of their predictive relations; (c) whether observed predictive relations are mediated by text comprehension; and (d) whether these relations vary or are developmentally invariant across 3 years of writing development. Based on analyses of the performance of students in Grades 4 ( $n = 246$ ), 5 ( $n = 242$ ), and 6 ( $n = 261$ ), the results supported morphological awareness, syntactic processing, and working memory as distinct yet correlated abilities that made independent contributions to predicting Chinese written composition, with working memory as the strongest predictor. However, predictive relations were mediated by text comprehension. The final model accounted for approximately 75% of the variance in Chinese written composition. The results were largely developmentally invariant across the 3 grades from which participants were drawn.

**Keywords:** Chinese children's written composition, text comprehension, mediation, working memory, morphological and syntactic processing

Although humans have been engaged in writing from the time they first began to read, considerably more research has been devoted to the study of reading compared with writing (Wagner et

al., 2011). In the late 19th century, studies of reading were relatively common while scientific studies of writing were just beginning to appear sporadically (Bazerman, 2008). In the last couple of decades, a great deal of writing research has been reported (for reviews, see Berninger & Chanquoy, 2012; Graham & Harris, 2009; Grigorenko, Mambrino, & Preiss, 2012; MacArthur, Graham, & Fitzgerald, 2006). However, with the exception of a relatively small literature that specifically addresses relations between reading and writing, research on writing and its development has proceeded largely independent of research on reading (Fitzgerald & Shanahan, 2000). In addition, the vast majority of studies on writing are limited to alphabetic writing systems. Finally, more research has been devoted to lower levels of reading and writing (i.e., decoding and spelling) compared to higher levels (i.e., comprehension and composition).

## Origins of Individual and Developmental Differences in Writing

If one asks children to produce written compositions, two empirical facts are immediately obvious. First, within a grade or restricted age range, individual differences are pronounced. Some children write fluently, producing longer, complex, and relatively error-free passages. Others write haltingly and are only able to

---

This article was published Online First March 3, 2014.

Connie Qun Guan, School of Foreign Languages, University of Science and Technology Beijing; Feifei Ye, School of Education, University of Pittsburgh; Richard K. Wagner, Psychology Department, Florida State University; Wanjin Meng, Psychology and Special Education Department, China National Institute of Education Sciences; Che Kan Leong, Department of Educational Psychology and Special Education, University of Saskatchewan.

This research was supported by National Office for Education Sciences Planning Grant DBA120179 and Engineering Research Institute Foundations of USTB Grant YJ2012-019 to Connie Qun Guan, National Institute of Education Science Grant GY2012013 to Wanjin Meng, and National Institute of Child Health and Human Development Grant P50HD052120 to Richard K. Wagner.

Correspondence concerning this article should be addressed to either Che Kan Leong, Department of Educational Psychology and Special Education, University of Saskatchewan, 28 Campus Drive, Saskatoon, SK, Canada S7N 0X1, or Wanjin Meng, Department of Psychology and Special Education, China National Institute of Education Sciences, Beijing, Bei-San-Huan-Zhong-Lu, #46, China 100088. E-mail: chekan.leong@usask.ca or wanjinmeng@yahoo.com



produce short passages replete with spelling and grammatical errors. The second obvious empirical fact is that developmental differences are obvious in writing samples produced by children from different grades.

The first model of writing to gain acceptance was proposed by Hayes and Flower (1980). According to the model, writing consisted of three parts: planning what you wanted to say, translating your ideas to print, and reviewing what you are writing. The model did not address individual or developmental differences, but a revision of the model did so indirectly by incorporating cognitive processes, such as working memory, that supported writing (Hayes, 1996). Individual and developmental differences in these supporting cognitive processes presumably would affect writing.

Several theories of writing were proposed subsequently that directly account for individual and developmental differences. Based on an analogy to the simple view of reading that explains individual and developmental differences in reading comprehension as the interaction between listening comprehension skills and decoding skills (Gough & Tunmer, 1986; Hoover & Gough, 1990), Juel, Griffith, and Gough (1986) proposed a simple view of writing in which individual and developmental differences in writing are accounted for by an interaction between quality of ideas and spelling ability. More recent theories of writing have been expanded to reflect the fact that writing operates under cognitive constraints such as limited working memory that presumably also affect reading comprehension as opposed to being uniquely related to writing (Berninger & Winn, 2006; Torrance & Galbraith, 2006).

### Relations Between Writing and Reading

Although research and pedagogy have viewed reading and writing as separate domains (Shanahan, 2006), when studies have measured both reading and writing the results suggest that reading and writing are closely related (Abbott & Berninger, 1993; Berninger, Abbott, Abbott, Graham, & Richards, 2002; Fitzgerald & Shanahan, 2000; Graham & Harris, 2000; Jenkins, Johnson, & Hileman, 2004; Juel, 1988; Juel et al., 1986; Shanahan, 1984; Tierney & Shanahan, 1991). Correlational analyses of measures of reading and writing indicate that approximately 50% of their variance is shared. When multiple indicators are available and latent variables can be used to reduce the influence of measurement error, up to 65% of the variance in reading and writing appears to be shared (Berninger, Abbott, et al., 2002; Shanahan, 2006).

It is not surprising that reading and writing are highly related. Writing and reading draw on analogous mental processes and knowledge, including (a) declarative knowledge (e.g., lexical knowledge of phonemic, graphemic and morphological awareness, syntax and text format); (b) procedural knowledge, such as accessing and using general knowledge to integrate various linguistic and cognitive processes; (c) domain knowledge, such as vocabulary, semantics and prior knowledge; and (d) metaknowledge or pragmatics in knowing the interactions of readers and writers and in monitoring one's own knowledge in composing and reading (Fitzgerald & Shanahan, 2000; Foorman, Arndt, & Crawford, 2011). Knowledge about reading might also be applied directly to writing or vice versa. Shanahan and Lomax (1986) compared models specifying reading-to-writing, writing-to-reading, and interactive relations in a study of second- and fifth-grade students. The reading-to-writing model was superior to the writing-to-

reading model, and the interactive model was superior to the reading-to-writing model for second grade. A recent study that modeled the codevelopment of reading and writing at the word, sentence, and passage level using latent change score modeling found support for a reading-to-writing model at the word and passage levels and for an interactive model at the sentence level (Ahmed, Wagner, & Lopez, in press).

Although reading and writing have much in common, there also are important differences. Reading involves recognition of words, whereas writing requires recall as well as spelling. Reading involves recognizing the grammatical structure of a sentence written by an author, whereas writing requires generating one's own sentence structures. Finally, reading requires following the arguments and organizational structure used by an author writing passages, whereas writing requires planning and designing argument structures and organizing sentences into coherent paragraphs and paragraphs into coherent documents (McCutchen, 2006). Given these differences, it is not surprising that writing is more difficult than reading.

### Relations Between English and Chinese Writing Systems

There are obvious differences between the English and Chinese writing systems but less obvious yet equally important similarities. Beginning with the most obvious difference, written English is a morphophonemic alphabet in which an orthography consisting of 26 letters as well as additional numbers and punctuation marks is used to represent all possible words. English is morphophonemic in that spellings represent pronunciation but with deviations that sometimes are attributable to meaning. In contrast, the character set of Chinese approaches 60,000 separate characters. Each character represents a spoken syllable that is a morpheme and often a word. Many of the 60,000 characters are low frequency, representing proper names or archaic words, and one can write or read 99% of modern Chinese with 2,400 characters (Schmandt-Besserat & Erard, 2008). But this still represents a difference of about two orders of magnitude compared to the number of letters that must be learned to write English. Grammar, syntax, and punctuation are often ambiguous and free-flowing in Chinese writing (Yan et al., 2012). Spoken Chinese is much more homophonic than is spoken English. Consequently, a large number of characters refers to the same syllable, and it is not possible to determine which of many meanings is intended without considering the surrounding context (Tan, Spinks, Eden, Perfetti, & Siok, 2005).

Turning to similarities, English and Chinese are both writing systems that convey information about pronunciation and meaning. Although it is commonly thought that Chinese characters are largely pictorial representations of concepts absent of pronunciation, approximately 90% of Chinese characters include a graphic element that indicates pronunciation, along with another graphic element that indicates meaning (Schmandt-Besserat & Erard, 2008). Similarity between writing in Chinese and English is suggested by a study of underlying dimensions of written composition in 160 Grade 4 and 180 Grade 7 Chinese children (Guan, Ye, Wagner, & Meng, 2013). They tested the generalizability of a five-factor model of writing developed by Wagner et al. (2011) from an analysis of English writing samples to Chinese writing samples. They asked the children to write two compositions and used the Systematic Analysis of Language Transcripts (SALT) program (Miller & Chapman, 2001) to code and



analyze the data. Guan et al. found that the five-factor model of macro organization, complexity, productivity, spelling and pronunciation, and handwriting fluency that was derived from English writing samples applied equally well to both fourth- and seventh-grade Chinese writers. The fourth- and seventh-grade writers differed in the latent means of the factors but not in the pattern of relations among factors.

Given both important similarities and differences between English and Chinese writing systems, it is difficult to predict in advance which aspects of knowledge about writing learned from the large number of studies of English writing will generalize to writing Chinese. Additional studies of Chinese writing will be necessary if we are to develop a theoretical framework for differentiating aspects of writing that are relatively language general and those that are relatively language specific.

### Individual and Developmental Differences in Chinese Writing

Given their role in English composition and differences between the English and Chinese writing systems, three potentially important predictors of Chinese writing are morphological awareness, syntax, and working memory.

**Morphological awareness.** Morphology is concerned with intraword and interword relations. Morphological awareness has been shown to play an important role in reading comprehension, particularly after controlling for word reading (Kirby et al., 2012; Kuo & Anderson, 2006). Morphological awareness is typically measured by tasks such as (a) morpheme discrimination in sorting out the odd item in orally presented four two-morpheme words (Packard et al., 2006), (b) morpheme production in producing a two-morpheme word with meaning identical to a target morpheme and another word with meaning unrelated to the target (Shu, McBride-Chang, Wu, & Liu, 2006), (c) morpheme transfer of homophonic two-character morphemes (Packard et al., 2006), and (d) morpheme analogy in generalizing a morphological relation from a pair of words to another pair by analogy (Kirby et al., 2012; Liu & McBride-Chang, 2010).

Because the Chinese language includes many homophones, morphological awareness is of particular importance to Chinese reading and writing (Hao, Chen, Dronjic, Shu, & Anderson, 2013; Kuo & Anderson, 2006; Liu & McBride-Chang, 2010; Packard et al., 2006; Shu et al., 2006; Zhang et al., 2012). Chinese morphology is predominantly that of morphological compounding. A compound can be defined as a word consisting of two or more words that are subjected to certain phonological and morphographic processes (Fabb, 1998). Chinese children have been shown to have a better developed sense of compounding than their American counterparts (Zhang et al., 2012). Semantic relatedness and types of morphemes in Chinese play different roles at different stages of reading literacy development in Chinese children (Hao et al., 2013). More proficient Chinese language users, compared with less proficient ones, have been shown to generate more two-character compound words from left-headed or right-headed base forms (Leong & Ho, 2008). Examples are 乐观 (optimistic) and 乐器 (musical instrument) from the base form of 乐. Because creating sentences in Chinese demands choosing appropriate characters and surrounding context to permit the reader to infer the correct morpheme, morphological awareness may be critically important to effective Chinese writing.

**Syntactic processing.** Even though many Chinese sentences are basically of the subject-verb-object (SVO) type, syntax is less straightforward in Chinese compared to English. As an example, the subject in a sentence may not always be expressed. The following simple sentence begins with the verb “downed” in 下雨了 (“Downed rain already” or “It rained”). As yet another example, the semimorphological marker 被 [bei] is meant to express unhappy or unexpected events. It is correct to say, 我們被 [bei]人打了 (“We are [were] beaten by others”), but it is anomalous to use the negation, \*我們被 [bei]人不打了 (“We were not beaten by others”).

Studies relating syntactic processing in Chinese to literacy acquisition are sparse. Yeung et al. (2011) used oral cloze task of the kind “My favorite food is \_\_\_\_\_.” to gauge syntactic skill. But this is more of a sentence completion task rather than a direct test of syntactic processing. Chik et al. (2012) included several measures of syntactic processing in a study of reading comprehension in Grades 1 and 2 Chinese children. In a hierarchical multiple regression equation, age, IQ and Chinese word reading accounted for 64% of the individual variation while composite syntactic skills added a significant 4% of the variation.

**Working memory.** Working memory is believed to be a key predictor of written composition because it provides the cognitive workspace in which writing processes are carried out (Abbott, Berninger, & Fayol, 2010; Berninger & Winn, 2006; Hayes, 1996, 2006; Hoskyn & Swanson, 2003; Kellogg, 1996, 1999, 2001, 2004; Kellogg, Whiteford, Turner, Cahill, & Mertens, 2013; McCutchen, 2000, 2011; Swanson & Berninger, 1996; Torrance & Galbraith, 2006; Vanderberg & Swanson, 2007). For example, Swanson and Berninger (1996) found working memory to be significantly correlated with writing after partialling out word knowledge. In particular, children with high memory span may allocate more resources to generating text rather than to transcription processes such as handwriting and spelling. Abbott et al. (2010) showed consistent and significant relations from spelling to text composing in their 5-year longitudinal study, relations that were explained using the construct of working memory. Children with strong spelling skills required fewer memory resources to translate ideas into written words and compositions than did children with weak spelling skills; more working memory resources were available to strong spellers relative to weak spellers to be applied to higher level aspects of writing.

Working memory is involved in transcribing and editing during writing as shown in a study by Hayes and Chenoweth (2006), who used articulatory suppression to place an additional load on working memory in a study of college undergraduates. The results were that participants in the articulatory suppression condition wrote more slowly and make significantly more errors compared to participants in a control condition.

### Predicting Chinese Writing at the Latent Construct Level

In general, there is a paucity of research on individual and developmental differences in Chinese writing. Most of the previous studies on predictors of Chinese writing have largely focused on relatively lower level skills such as character writing quality (Bi, Han, & Zhang, 2009; Guan, Liu, Ye, Chan, & Perfetti, 2011; Guan et al., 2013; Perfetti & Guan, 2012; Tan et al., 2005). For



example, Tan et al. (2005) examined relations between reading and writing Chinese characters for groups of beginning and intermediate readers. Partial correlations between reading and writing after controlling for nonverbal intelligence were .50 ( $p < .001$ ) and .47 ( $p < .001$ ) for beginning and intermediate readers, respectively. However, the cross-sectional and correlational design of the study precluded determining the directionality of these relations (Bi et al., 2009).

More recently, Yan et al. (2012) reported a longitudinal study of writing at the passage level as opposed to the level of the individual character. In their study, the writing quality of 9-year-old Chinese students was predicted by earlier measures of vocabulary knowledge, Chinese word dictation, phonological awareness, speed of processing, speeded naming, and handwriting fluency were all significantly associated with writing, after controlling for age.

A limitation of the studies of predictors of Chinese writing just described, as well as many studies of English writing, is that the constructs were represented by single observed variables as opposed to latent variables with multiple indicators. When constructs are represented by single observed variables, the obtained correlation and regression coefficients are affected by measurement error and method variance. One consequence is that measurement error and method variance can make it appear as though the constructs are distinct from one another, when in fact, they all are measuring an identical underlying construct such as language or verbal aptitude. Conversely, when constructs are represented by latent variables with multiple indicator, the effects of measurement error and method variance can be reduced or eliminated depending on the design of the study.

Several studies have begun to look at predictors of Chinese writing at the latent variable level. For example, in a study of component processes in language literacy in 361 15-year-old Chinese students, Leong and Ho (2008) used stimulus cartoon pictures to elicit students' essay writing. They also obtained measures of morphological processing, character and word correction, text segmentation, dictation, copying words and text, text comprehension, oral reading and reading fluency. Exploratory factor analysis was used to examine underlying dimensions of task performance. Six components accounted for 67% of the total variance, with half of the variance accounted for by the component of lexical knowledge. This consisted of morphological processing, correct usage of lexical items, segmentation of text passages and writing to dictation. These patterns were largely validated in a confirmatory factor analysis with a new group of 1,164 15-year-old Chinese students (Leong, Ho, Chang, & Hau, 2013). The strongest correlations among factors were obtained for correlations between the reading and writing factors.

### The Present Study

The goal of the present study was to test opposing views about four issues concerning predictors of individual differences in Chinese written composition: (a) Whether morphological awareness, syntactic processing, and working memory represent distinct and measureable constructs or are manifestations of general language ability; (b) whether they are important predictors of Chinese written composition, and if so, the relative magnitudes and independence of their predictive relations; (c) whether observed predictive

relations are mediated by text comprehension; and (d) whether these relations vary or are developmentally invariant across 3 years of writing development.

1. *Distinct and measureable constructs in Chinese or just manifestations of general language ability?* Based on the existing literature of predictors of writing in English, and on the nature of the Chinese writing system, morphological awareness, syntax, and working memory are potentially important predictors of Chinese written composition. However, because previous studies typically have included only one of these constructs, and the constructs have been represented as single indicator observed variables, it remains important to determine whether these are meaningful distinct and measureable constructs as opposed to simply measures of general language ability. This issue was addressed in the present study by measuring each construct with multiple indicators and then using confirmatory factor analysis to test alternative models. One of these posited a three-factor model with morphological awareness, syntactic processing, and working memory as distinct yet potentially correlated abilities; another posited a single-factor model representing general language ability.

2. *Important predictors of Chinese written composition, and if so, what are the relative magnitudes and independence of their contributions to prediction?* Although there is a theoretical rationale for expecting morphological awareness, syntactic processing, and working memory to be important predictors of Chinese written composition, the empirical evidence is scant. Only two studies have used morphological processing as a predictor of writing in Chinese (Leong & Ho, 2008; Leong et al., 2013). No studies have used syntax or working memory as a predictor of Chinese writing. By including all three constructs as predictors in the present study, it was possible to determine (a) whether each was an important predictor of Chinese written composition, (b) the relative magnitudes of their predictive relations, and (c) whether their contributions to predictions were independent or redundant. Bivariate relations between latent variables representing each of the three predictors and the criterion were used to determine whether each was an important predictor of Chinese written composition. We used dominance analysis (Azen & Budescu, 2003; Budescu, 1993) to compare the relative magnitude of these predictive relations. We used structural equation models that included all three constructs as simultaneous predictors to determine whether their contributions to prediction were independent or redundant. It would be possible for the constructs to be distinct, yet for their predictive relations to Chinese written expression to be redundant if their prediction of writing was due to their correlation with general language ability and language ability in turn predicted writing. Alternatively, each construct could be capturing different aspects of language that were independently related to writing.

3. *Are any observed predictive relations mediated by text comprehension?* Because of similarities and differences between reading and writing, predictive relations between the three key constructs of morphological awareness, syntactic processing, and working memory and the dependent variable of Chinese written composition might be mediated by text comprehension. Individual and developmental differences in morphological awareness, syntactic processing, and working memory have been shown to be important predictors of reading, although much of this research is limited to reading alphabetic writing systems. However, because of the constructive nature of writing, they may be involved in

writing to a greater extent than they are in reading. In the present study, we compared alternative models that proposed that predictive relations between morphological awareness, syntactic processing, and working memory were (a) unmediated, (b) partially mediated, or (c) fully mediated by text comprehension.

A variable is a mediator “to the extent that it accounts for the relationship between the predictor and the criterion” (Baron & Kenny, 1986, p. 1176). According to Baron and Kenny (1986), three conditions must be met to establish M as a mediator of the predictive relation between X and Y: (a) X must significantly predict Y, (b) X must significantly predict M, and (c) M must significantly predict Y controlling for X. Complete mediation is said to occur when the direct effect of X on Y decreases to zero with the addition of potential mediator M. Partial mediation is said to occur when the direct effect of X on Y decreases nontrivially but not to zero with the addition of potential mediator M. No mediation is said to occur when the direct effect of X on Y is substantially unchanged with the addition of potential mediator M.

To test for mediation, we used the two models represented in Figure 1. Figure 1a depicts a structural equation model that specifies morphological awareness, syntactic processing, and working memory as predictors of writing. Fitting this model to the data provides estimates of the unique contributions to prediction of writing made by the three constructs. Figure 1b depicts a structural equation model in which text processing has been added as a mediating latent variable. The mediating relations are represented by the indirect effects from each of the three predictors through text comprehension to writing. The magnitude of the direct effects from the three predictors to writing in Figure 1b after the mediator variable of text comprehension is added determines whether there is evidence for full, partial, or no mediation. Full mediation would be indicated if the direct effects are no longer significantly greater than zero. Partial mediation would be indicated if the direct effects are significantly less than they were in the unmediated model but remain significantly greater than zero. No mediation would be indicated if there are no significant differences between the magnitudes of the direct effects for the mediated and unmediated models.

4. *Developmental differences or invariance?* In a previous study of the underlying dimensions of Chinese written composition, a five-factor model of individual differences in written composition that originally was developed from analyses of English writing

samples produced by first- and fourth-grade students was found to generalize to Chinese writing samples produced by fourth- and seventh-grade students (Guan et al., 2013; Wagner et al., 2011). These studies suggest surprising consistency in the underlying dimensions of written composition across grade and language. However, neither of these studies examined predictors of written composition. As such, they are not informative about whether predictive relations between morphological awareness, syntactic processing, and working memory vary developmentally or are relatively invariant. In the present study, we analyzed the data separately by grade to examine invariance in our measurement model and in relations among constructs across the developmental range represented by fourth through sixth grades.

The present study differs from a recently published article by Guan, Ye, Meng, and Leong (2013), which drew on a subset of poor readers and writers from the same large data pool. That study examined the transactional process of Chinese reading–writing difficulties and used similar cognitive and linguistic tasks. That study showed from hierarchical multiple regression analyses that verbal working memory contributed to individual variation in written composition by poor text comprehenders but not good readers. The study also provided insight into the quality of the students’ writing from a qualitative analysis of some sample written compositions. As discussed in this section, the emphasis of the present article was on the relative magnitude of predicting individual differences in Chinese written composition by the constructs of morphological awareness, syntactic processing and working memory and on the predictive relations mediated by text comprehension.

## Method

### Participants

A total of 749 students took part in the study. They were recruited to participate in a larger longitudinal study about assessment and intervention of writing disabilities (Grant DBA120179) and Chinese writing model (Grant YJ2012-019) conducted by the first author. These participants were drawn from Grades 4, 5, and 6 in one primary school in Ningbo, Zhejiang Province, China. According to the municipal educational bureau report, the students’ parents’ average annual salary was about 25,000 USD; their

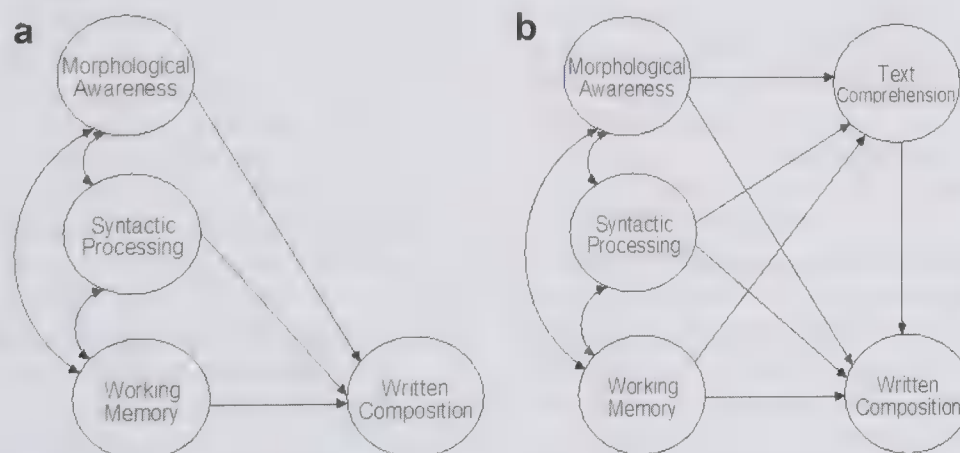


Figure 1. Proposed models of prediction of written composition.



demographic information and social economic status are representative of the middle class in China (NIES, 2012). Their parents signed the informed consent form before their children actually participated in this study. There were 246 Grade 4 students from six classes ( $n_{\text{boy}} = 142$ ,  $n_{\text{girl}} = 104$ ,  $M_{\text{age}} = 9.76$  years,  $SD_{\text{age}} = 0.84$ ), 242 Grade 5 students from six classes ( $n_{\text{boy}} = 129$ ,  $n_{\text{girl}} = 113$ ,  $M_{\text{age}} = 11.01$  years,  $SD_{\text{age}} = 0.84$ ), 261 Grade 6 students from six classes ( $n_{\text{boy}} = 155$ ,  $n_{\text{girl}} = 106$ ,  $M_{\text{age}} = 12.31$  years,  $SD_{\text{age}} = 0.70$ ). For the total sample, the mean age was 11.05 years ( $SD = 1.32$  years). To assess the possible dependence among the students within a class, we calculated the intraclass correlation for the three writing variables and found the nesting effect was minimal (intraclass correlations [ICCs]  $< .047$ ). Thus, we proceeded with the analysis by treating these students as independent.

The group tasks were administered in the classrooms of the students. Three full-time and nine part-time Chinese-speaking research assistants were given several days' intensive training on the rationale of the project, the reasons and designs of all the tasks, and specifics of administration before their field work in the school. These experienced assistants were carefully supervised by the first author to ensure high fidelity of data collection.

## Tasks and Procedure

Multiple indicators were obtained to provide latent variables representing five constructs:

**Written composition (WC).** We asked the children to write three kinds of compositions: narration, argumentation, and exposition. These kinds of writing are representative of important writing tasks (Berman & Nir-Sagiv, 2007; Britton, 1994). Narratives focus on people, their action, motivation, and events unfolding in a temporal sequence; expository compositions focus on issues with ideas unfolding in logical structure. Different from narrations and expositions, argumentation compositions require writers to argue and counterargue, all based on plausibility and factual information (Reznitskaya, Anderson, & Kuo, 2007). These written compositions were scored according to the three aspects of expressiveness, content and commentary.

Narrative writing (WNar) was produced in response to four black-and-white line drawing cartoons without words and titles from Leong and Ho (2008). These cartoons have a universal appeal to all ages and can be interpreted flexibly from different perspectives. There were four basic elements in the cartoons: a boy reading while a girl is coming forward, a boy and girl in conversation, the two children having different opinions with an ensuing argument, the girl getting away and the boy falling. From these simple but integrated themes the students were asked to write short compositions from their personal experience to construct a text-base to describe the scenes and to express their meaning and emotion (see Kintsch & Kintsch, 2005). They should also provide appropriate discussion and a title. This task was given to groups of students in 20 min, and they were requested to write individually between 150 and 500 words. A total score of 100 was given to each student. Scoring was by two research assistants according to expressive aspects (40% of total score), content including title (40% of total score), and commentary (20% of total score). Expressive aspects included total number of words, total number of new words, word choice of low frequent vocabulary, and lexical density. The content included four aspects of topic, main idea, body,

and conclusion. Commentary included two aspects of objective discussion and subjective comments on the theme of writing. Any disagreement on scoring was reexamined by a third assistant and resolved accordingly. Interrater reliability of the original two scorers represented by the Pearson product-moment coefficient for the task was .85, and test-retest reliability was .75.

Argumentation writing (WArg) was on the advantages and disadvantages of watching television for elementary school children. Students were asked to state the pros and cons of watching television, and give reasons with examples to illustrate their points. They were also instructed to provide appropriate discussion for each point. Similar to expository writing, this task was administered to groups of students who were given 20 min to write compositions of between 150 to 500 words. The scoring procedure was the same as the narrative writing task. Interrater reliability for the task as a whole was .88, and test-retest reliability was .76.

Expository writing (WExp) was writing on the topic of "My Favorite Pet/Toy." Students were asked to name one of their favorite pets (or toys if there were no pets) and describe their detailed features and other interesting characteristics. This task was also given as a group-administered writing task in 20 min with students being requested to write individually between 150 and 500 words. The scoring procedure was the same as the other two writing tasks. Interrater reliability for the task as a whole was .78, and test-retest reliability was .86.

**Text comprehension (TC).** Eight short text passages were adapted from Leong, Tse, Loh, and Hau (2008) and rewritten in simplified Chinese characters. Four of the eight text passages were narrative pieces, and the other four were expository essays. These eight short essays were carefully balanced in syntactic complexity, ranged in length from six sentences to 13 sentences, and the contents were all of interest to children between the ages of 9 and 12 years. An example was the passage "Alfred Nobel," which gives an account of the contribution of the inventor and Nobel Prizes and contains eight sentences (two simple and six compound sentences). These eight passages were followed by three written open-ended comprehension questions each. The questions drew on higher order thinking, such as hypothesizing, using schemata, questioning, citing evidence, and verifying ideas and integrating them.

The text comprehension task was administered to the students as a class in 40 min plus 10 min for two short practice examples to explain the task. In one practice example with three sentences the translated text is as follows: "One cold winter day, a group of displaced persons arrived at the small town. They were ghastly pale and utterly exhausted. The people of the small town cooked them hot meals." In the first of the two questions the children were asked to discuss verbally what they would do if they were the displaced persons and given free food. The whole class was asked to give the "best" answers and was told these would be graded according to the depth of meaningfulness of the answers from a credit of 3, to 2, then 1 or 0 for an irrelevant or implausible answer. An answer such as "I will say 'thank you' and eat the food" was given a score of 1. An answer such as "I will say 'thank you' but also offer to do some work in return before taking food from strangers" was given a score of 3. The maximum score for the whole task was 72 (8 passages  $\times$  3  $\times$  3).

The principles of scoring the written answers were on the basis of problem solving and transforming knowledge and not merely



telling it (Bereiter & Scardamalia, 1987, p. 341), of explanatory and not just descriptive or factual answers, and of “envisionment” of text-worlds (Langer, 1986). The children were further told to work quickly and to concentrate on making meaning and not worry about sentence construction and spelling since they had to read the eight passages and to write the answers to all the 24 open-ended questions on the protocols in the span of 40 min.

To ensure consistency of grading, each set of written protocols was marked by two research assistants according to the grading principles explained above. Interrater reliability for the protocols for the eight passages as a whole was .91. This coefficient indicates that the eight passages as a whole and the answers to the comprehension questions were consistent and useable. The mainly narrative texts (Passages A1, 2, 3, and 4) based on genre and structure constituted one indicator text comprehension 1 (TC1). Cronbach alpha was .77. The mainly expository texts (Passages B 5, 6, 7, and 8) formed the second indicator text comprehension 2 (TC2). Interrater reliability for this TC2 was .78, and test–retest reliability was .81.

**Working memory (WM).** The working memory construct was represented by two tasks: a verbal span working memory task (VSWM) involving unrelated sentences and an operation span working memory (OSWM) task involving numbers and very simple Chinese words.

A verbal span working memory (VSWM) task was adapted from that used by Leong et al. (2008) and Leong and Ho (2008). It was based on the rationale and format of Daneman et al. (Daneman & Carpenter, 1980; Daneman & Merikle, 1996), as modified by Swanson (1992). Six sets of two, three and four sentences, all unrelated in meaning, were read orally by the experimenter to groups of students. They first listened to each set of two, three, or four sentences, plus a comprehension question, all spoken in Putonghua, and were then to write down on designated forms their short answers to the comprehension question and the last word in each sentence of the set. The total testing time for this task was 20 min, and all the answers were scored independently by two research assistants. One point was awarded for each correct answer and the maximum score was 24. Interrater reliability for the task was .83, and test–retest reliability was .77.

An operation span working memory (OSWM) task was modeled after the operation span task of Engle, Tuholski, Laughlin, and Conway (1999). Groups of students heard six sets of three or four sentences, each of which involved very simple mental arithmetic calculation with either a correct (YES) or wrong (NO) answer, followed by a simple spoken word. Students had to wait until the end of each sentence set before writing down on the designated forms just YES/NO to the answers of the simple calculation and the one word at the end in the correct order. An example of a three-sentence set is as follows: “Is  $16 - 9 = 7$ ? (Bear) YES/NO; Is  $12 \times 2 = 24$ ? (Bus) YES/NO; Is  $20 - 6 = 12$ ? (Book) YES/NO.” The total testing time for this task was 15 min, and the maximum score was 21. Interrater reliability for the task was .98, and test–retest reliability was .76.

**Morphological awareness (Morph/MP).** There were two indicators for this construct: a morphological compounding (MorCom/Morph1) task from Leong and Ho (2008) and a morphological chain (MorCha/Morph2) task.

A morphological compounding (MorCom) task contained two parts that varied in generating left-headed or right-headed two-

character morphological compound words with eight base items each for a total group administration time of 12 min. Students could freely choose any five base forms to produce as many “right-headed” two-character words in the available time and any six base forms to produce as many “left-headed” two-character words in the total time of 6 min allotted. Two research assistants scored the freely affixed items to the base forms. Interrater reliability was .83, and test–retest reliability was .79. The Cronbach Alpha internal consistency reliability of all the items for this measure was .70.

A morphological chain (MorCha) task required the participants to provide as many different two-character compound words from the left-headed base character as possible in 5 min time. The constraint was that the same base form and homophonic base forms could not be repeated. Two research assistants scored the freely affixed items to the base forms. Interrater reliability was .98, and test–retest reliability was .82. Cronbach Alpha was .74.

**Syntactic processing (SP).** Syntactic processing plays an important role in helping language users to understand the appropriate relationship between topics and comments and the interpretation of the sentence. A topic is what the sentence is about, and the comment is the rest of the sentence, separable from the topic by a pausal marker. A topic sets a “spatial, temporal, or individual framework within which the main prediction holds” (Li & Thompson, 1989, p. 85). Syntactic processing is thus an interactive process with lexical knowledge and sentential context mutually influencing each other. There were two tasks: syntax construction and syntax integrity.

A syntax construction (SynCon) task consisted of 10 scrambled sentences scrambling mostly two-character words. Students were asked to recombine the words in the scrambled sentences to come up with the correct sequence of the lexical items to make the sentences grammatically correct in the recombination. The administration time was 20 min. All the answers were scored by two research assistants. Interrater reliability was .88, and test–retest reliability was .80. Cronbach alpha of the internal consistency of all the items for this measure was .71.

A syntax integrity (SynInt) task required error detection and correction. The syntax integrity task assessed the students’ understanding and correct usage of syntactic structure. The students were asked to read each of the 20 short grammatically anomalous sentences, detect the error in the syntactic pattern and to correct that error. There were 20 sentences, and the testing time was 25 min. All the answers were scored by two research assistants. Interrater reliability was .92, and test–retest reliability was .81. Cronbach alpha of the internal consistency of all the items for this measure was .82.

## Procedures

The tasks were administered to groups of students over three consecutive days. The verbal span working memory task, morphological compounding task, text comprehension 1, and narrative writing task were administered on Day 1. The syntax construction task, text comprehension 2, and argumentation writing task were administered on Day 2. The syntax integrity task, the morphological chain task, the operation span working memory task, and expository writing task were administered on Day 3. Instructions



for each task were audio-taped and played to the students groups, so that all the tasks were administered uniformly across groups.

## Data Analysis

The data analysis was carried out in three stages after data screening as follows.

**Confirmatory factor analysis.** The first stage was to assess the construct validity and measurement invariance of the proposed latent variables. At this stage, we first conducted confirmatory factor analysis (CFA) for each of Grades 4, 5, and 6. In each CFA model, one of the factor loadings for each factor was fixed to be one for scale dependency in model identification. In the second step, we assessed measurement invariance across grades. The purpose of testing measurement invariance was to establish that either partial- or full-measurement invariance was established across grades. Failing to do so would preclude meaningful comparisons across grades because of concern that the latent variables were not comparable.

There are several forms of invariance in the procedure. Here we tested metric invariance (equal factor loadings) and scalar invariance (equal intercepts) using multigroup CFAs. Metric invariance is required for comparing latent means, while there is debate on whether scalar invariance is needed (Ployhart & Oswald, 2004). A stepwise procedure was adopted to assess measurement invariance (Vandenberg & Lance, 2000): (a) A baseline model was analyzed without any equality constraints for corresponding factors; (b) an equal factor loading model was analyzed with equality constraints imposed on corresponding factor loadings (metric invariance). If all factor loadings were invariant, we continued to (c) assess invariance of intercept (scalar invariance). If all factor loadings were not invariant, we found out which variables had equal factor loadings and then among these variables, which had equal intercepts. The chi-square difference test was used to assess the invariance of factor loadings and intercepts. Chi-square difference testing was conducted using the Satorra-Bentler adjusted chi-square (Satorra, 2000; Satorra & Bentler, 2001). With measurement invariance established, latent means were compared across grades with latent standardized effect sizes reported (Choi, Fan, & Hancock, 2009).

**Structural equation models.** The second stage of data analysis consisted of testing alternative structural models to estimate the strength of predictive relations between morphological awareness, syntactic processing, and working memory as predictors of written composition and the potential mediating effects of text comprehension on these predictive relations. Chi-square difference testing was used to compare results across grades.

For the CFA and structural equation modeling (SEM) analyses, the goodness of fit between the data and the specified models was estimated by employing the comparative fit index (CFI; Bentler, 1990), the Tucker-Lewis index (TLI; Bentler & Bonett, 1980), the root-mean-square error of approximation (RMSEA; Browne & Cudeck, 1993), and the standardized root-mean-square residual (SRMR; Bentler, 1995). CFI and TLI guidelines of greater than 0.95 were employed as standards of good fitting models (Hu & Bentler, 1999). Different criteria are available for RMSEA. Hu and Bentler (1995) used .06 as the cutoff for a good fit. Browne and Cudeck (1993) and MacCallum, Browne, and Sugawara (1996) presented guidelines for assessing model fit with RMSEA: values

less than .05 indicate close fit, values ranging from .05 to .08 indicate fair fit, values from .08 to .10 indicate mediocre fit, and values greater than .10 indicate poor fit. A confidence interval of RMSEA provides information regarding the precision of RMSEA point estimates and was also employed as suggested by MacCallum et al. A SRMR < .08 indicates a good fit (Hu & Bentler, 1999). All CFA and SEM analyses were performed with Mplus 6.1 (Muthén & Muthén, 2010).

**Dominance analysis.** The third stage of data analysis consisted of dominance analysis (Azen & Budescu, 2003) to assess the unique contribution and relative importance of morphological awareness, syntactic processing, and working memory in accounting for variance in written composition. For the dominance analysis, the dependent variable written composition was calculated as the sum of standardized scores of the three writing tasks. The predictors, working memory, morphological processing, and syntactic processing were also calculated as the sum of the standardized scores of the corresponding tasks.

The purpose of dominance analysis is to address the problem that the relative importance of correlated predictors is affected by the other predictors included in or excluded from the model (Cohen, Cohen, West, & Aiken, 2003; Courville & Thompson, 2001). Common measures of relative importance, including standardized regression coefficient, zero-order correlation, partial correlation, semipartial correlation, are affected by this phenomenon. More recently, dominance analysis, developed by Budescu (1993) and refined and extended by Azen and Budescu (2003), presents a better alternative for analysis of predictor importance and provides a general approach to measure relative importance in a pairwise fashion in the context of all models that contain some subsets of the other predictors (Azen & Budescu, 2003). Dominance analysis is able to answer the key question of predictor importance: "Is variable  $X_i$  more or less (or equally) important than variable  $X_j$  in predicting  $Y$  in the context of the predictors included in the selected model?" (Azen & Budescu, 2003, p. 145).

Several measures of dominance were introduced that differ in the strictness of the dominance definition. Here we adopted the strictest definition of dominance, complete dominance. We illustrate this with three predictors ( $X_1$ ,  $X_2$ ,  $X_3$ ), as in the current study, to predict one criterion variable. All possible model combinations of predictors were examined, including three subset models with only one predictor, three models with two predictors, and one model with all four predictors, resulting in a total of seven subset models. Predictor  $X_1$  is said to have complete dominance over predictor  $X_2$  when unique variance contribution of Predictor  $X_1$  is greater than Predictor  $X_2$  in each of the subset models to which both  $X_1$  and  $X_2$  could make additional contribution, i.e., the null model without any predictor, and the model with  $X_3$ .

To generalize dominance results beyond the studied sample, we followed Azen and Budescu (2003) in calculating the standard error of dominance across repeated sampling and the reproducibility of the present dominance in the population. Let  $D_{ij}$  denote a measure of dominance, which equals 1 if  $X_i$  dominates  $X_j$ , equals 0 if  $X_j$  dominates  $X_i$ , and equals .5 if dominance cannot be established between the two predictors. A distribution of  $D_{ij}$  could be simulated by obtaining this measure over many (e.g., 1,000) repeated samples with replacement, which are generated using the

bootstrap procedure. The average of these dominance values over all bootstrap samples,  $\bar{D}_{ij}$ , represents the expected level of dominance of  $X_i$  over  $X_j$  in the population. The standard error of  $D_{ij}$ ,  $SE(D_{ij})$ , is the standard deviation of  $D_{ij}$  over all bootstrap samples.  $\bar{D}_{ij}$  closer to 0 or 1 indicates a strong case for a clear directional dominance, while that close to .5 suggests indeterminacy of dominance. The percentage of the bootstrap samples that replicates a dominance of, for example,  $X_i$  over  $X_j$ , in the studied sample is termed reproducibility, which states the probability that  $X_i$  dominates  $X_j$  and determines a confidence level on that probability.

## Results

### Preliminary Analyses

Table 1 displays the means, standard deviations, skewness, and kurtosis of the various measures used in the study by grade level. In addition, the Shapiro-Wilk test was used to examine the normality of these measures. Several variables were not normally distributed, and there were moderate ceiling effects for operational span working memory. The assumption of multi-

variate normality was violated, Mardia's skewness = 75.04,  $p < .001$ , and Mardia's kurtosis = 156.65,  $p < .001$ . To address the nonnormality, Satorra-Bentler correction was implemented for both model fit and parameter estimation by using maximum likelihood with robust standard errors (MLR) estimation. Table 2 presents the intercorrelations of these measures by grade.

### Confirmatory Factor Analyses

The results of confirmatory factor analyses carried out by grade using the 11 tasks as indicators of the latent variables indicated that the five latent factors were measured well with these tasks for each grade. Specifically, the fit of the confirmatory factor model was satisfactory for Grade 4,  $\chi^2(34) = 90.57$ ,  $p < .001$ , RMSEA = .08 with 90% confidence interval (.06, .10), CFI = .96, TLI = .93, SRMR = .04; satisfactory for Grade 5,  $\chi^2(34) = 79.54$ ,  $p < .001$ , RMSEA = .08 with 90% confidence interval (.06, .10), CFI = .96, TLI = .94, SRMR = .03; and satisfactory for Grade 6,  $\chi^2(34) = 94.62$ ,  $p < .001$ , RMSEA = .08 with 90% confidence interval (.06, .10), CFI = .96, TLI = .93, SRMR = .04. Table 3 presents the standardized factor loadings and the correlations among factors. The correlations between the three predictors of morphological aware-

Table 1  
*Descriptive Statistics of All Tasks by Grade*

Task	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	Shapiro-Wilk	<i>p</i>
Grade 4 ( <i>n</i> = 246)						
Verbal Span Working Memory	15.09	6.94	-0.25	-1.30	0.92	.000
Operational Span Working Memory	16.30	5.13	-1.28	0.70	0.82	.000
Morphological Compounding	10.05	3.60	0.47	-0.05	0.97	.000
Morphological Chain	10.09	4.92	0.43	-0.79	0.95	.000
Syntax Construction	11.38	3.63	-0.26	-0.76	0.97	.000
Syntax Integrity	11.50	3.71	0.15	-0.48	0.98	.003
Text Comprehension Task 1	14.29	6.84	0.77	0.65	0.95	.000
Text Comprehension Task 2	14.43	5.61	0.24	-0.54	0.98	.004
Narrative Writing	52.09	9.42	-0.69	0.53	0.92	.000
Argumentation Writing	47.49	11.27	-0.11	-1.44	0.98	.000
Expository Writing	51.92	9.63	-0.80	0.60	0.90	.000
Grade 5 ( <i>n</i> = 242)						
Verbal Span Working Memory	17.55	5.68	-0.91	-0.41	0.91	.000
Operational Span Working Memory	17.04	4.27	-1.78	3.16	0.80	.000
Morphological Compounding	12.35	4.30	-0.12	0.07	0.98	.008
Morphological Chain	12.82	5.39	0.18	-0.54	0.98	.001
Syntax Construction	13.74	3.62	-0.71	1.53	0.95	.000
Syntax Integrity	12.96	3.42	-0.70	0.65	0.96	.000
Text Comprehension Task 1	19.00	7.69	-0.28	-1.11	0.94	.000
Text Comprehension Task 2	19.24	7.32	-0.28	-0.86	0.97	.000
Narrative Writing	60.65	9.33	-0.37	-0.38	0.96	.000
Argumentation Writing	56.92	13.54	-0.32	-0.63	0.97	.000
Expository Writing	62.69	10.60	-1.33	1.66	0.86	.000
Grade 6 ( <i>n</i> = 261)						
Verbal Span Working Memory	19.31	5.38	-1.05	-0.35	0.83	.000
Operational Span Working Memory	18.83	3.74	-1.92	2.79	0.64	.000
Morphological Compounding	14.09	4.16	0.24	0.75	0.98	.004
Morphological Chain	14.25	6.61	0.06	-1.18	0.95	.000
Syntax Construction	15.44	2.48	-1.36	3.00	0.90	.000
Syntax Integrity	14.44	3.12	-0.88	1.21	0.94	.000
Text Comprehension Task 1	21.25	7.72	0.03	-0.89	0.97	.000
Text Comprehension Task 2	21.83	6.96	-0.40	-0.52	0.97	.000
Narrative Writing	75.73	13.99	-1.36	1.59	0.86	.000
Argumentation Writing	67.26	14.40	-0.85	0.42	0.94	.000
Expository Writing	70.78	12.92	-1.29	1.88	0.89	.000



Table 2  
Correlations of Tasks for Grades 4, 5, and 6

Variable	1	2	3	4	5	6	7	8	9	10	11
Grade 4 ( <i>n</i> = 246)											
1. Verbal Span Working Memory	—										
2. Operational Span Working Memory	.59	—									
3. Morphological Compounding	.41	.29	—								
4. Morphological Chain	.22	.19	.44	—							
5. Syntax Construction	.30	.27	.29	.28	—						
6. Syntax Integrity	.32	.23	.24	.29	.34	—					
7. Text Comprehension Task 1	.56	.44	.55	.45	.50	.41	—				
8. Text Comprehension Task 2	.60	.43	.61	.41	.45	.42	.79	—			
9. Narrative Writing	.49	.51	.53	.44	.50	.38	.82	.64	—		
10. Argumentation Writing	.48	.46	.51	.39	.43	.33	.72	.62	.79	—	
11. Expository Writing	.41	.48	.39	.34	.30	.22	.59	.44	.74	.66	—
Grade 5 ( <i>n</i> = 242)											
1. Verbal Span Working Memory	—										
2. Operational Span Working Memory	.59	—									
3. Morphological Compounding	.34	.37	—								
4. Morphological Chain	.22	.30	.47	—							
5. Syntax Construction	.34	.30	.28	.25	—						
6. Syntax Integrity	.24	.23	.28	.31	.58	—					
7. Text Comprehension Task 1	.55	.49	.50	.34	.51	.37	—				
8. Text Comprehension Task 2	.50	.47	.43	.32	.52	.38	.78	—			
9. Narrative Writing	.45	.40	.42	.39	.35	.23	.78	.61	—		
10. Argumentation Writing	.40	.35	.32	.27	.42	.35	.64	.48	.73	—	
11. Expository Writing	.47	.46	.44	.37	.46	.37	.64	.47	.69	.69	—
Grade 6 ( <i>n</i> = 261)											
1. Verbal Span Working Memory	—										
2. Operational Span Working Memory	.68	—									
3. Morphological Compounding	.45	.38	—								
4. Morphological Chain	.43	.37	.45	—							
5. Syntax Construction	.28	.32	.18	<b>.12</b>	—						
6. Syntax Integrity	.31	.33	.25	.17	.48	—					
7. Text Comprehension Task 1	.65	.54	.65	.38	.39	.37	—				
8. Text Comprehension Task 2	.66	.57	.55	.29	.51	.45	.82	—			
9. Narrative Writing	.57	.50	.51	.35	.40	.35	.69	.67	—		
10. Argumentation Writing	.58	.52	.44	.29	.33	.35	.68	.66	.72	—	
11. Expository Writing	.47	.42	.45	.22	.27	.28	.58	.59	.74	.62	—

Note. All coefficients are statistically significant at the .05 level except the one in bold.

ness, syntactic processing, and working memory and the criterion of written composition were substantial at each grade level, ranging from a low of .57 to a high of .79.

The adequate model fits and the moderate correlations between the three predictors of morphological awareness, syntactic processing, and working memory supported the view that these represent distinct and measurable abilities. The alternative view that they are just manifestations of a single underlying factor of general language ability was tested by a model that represented the indicators of morphological awareness, syntactic processing, and working memory as indicators of a single general language factor. This model resulted in a significantly poorer fit at each grade, with  $\Delta\chi^2$  values of 65.32, 134.27, and 88.68 for Grades 4, 5, and 6, respectively, all significant at  $p < .001$  for  $\Delta df = 7$ .

We examined the measurement invariance between grades using multigroup CFA (see Table 4). The baseline model resulted in a good fit. The model with equal loadings resulted in a significantly poorer fit. We examined each variable individually and found that narrative writing (WNAR) had a different loading for Grade 5. We further tested the invariance on intercepts and found that the model with equal intercepts resulted in

a significantly poorer fit. We examined each variable individually and found that operational span working memory (OSWM) had a different intercept for Grade 5, and text comprehension task 2 had a different intercept for Grades 5 and 6. These results suggest that partial measurement invariance held across grades.

Table 5 presents the latent means and variances of the five factors for each grade. The fifth graders had significantly higher means than the fourth graders, and the sixth graders had significant higher means than the fourth and fifth graders ( $ps < .001$ ). The latent standardized effect sizes (Choi et al., 2009) for pairwise comparisons on the latent means (as shown in Table 5) suggested that the mean difference was medium between Grades 4 and 5, small to medium between Grades 5 and 6, and large between Grades 4 and 6.

In summary, the results indicated that the latent variables were well measured by their indicators. Measurement invariance was largely supported, allowing us to compare latent means across grades. Sixth grade students had significantly higher means than the fifth-grader students, who in turn had significantly higher means than the fourth grade students.

Table 3  
Standardized Factor Loading (Standard Error) and Interfactor Correlations From Confirmatory Factor Analysis

Task	Grade 4					Grade 5					Grade 6				
	WM	MP	SP	TC	WC	WM	MP	SP	TC	WC	WM	MP	SP	TC	WC
Verbal Span Working Memory	.84 (.05)					.79 (.04)					.89 (.03)				
Operational Span Working Memory	.71 (.06)					.74 (.05)					.77 (.04)				
Morphological Compounding		.75 (.05)					.79 (.06)					.87 (.07)			
Morphological Chain		.58 (.05)					.60 (.06)					.51 (.06)			
Syntax Construction			.64 (.07)					.87 (.05)					.74 (.06)		
Syntax Integrity			.53 (.06)					.67 (.06)					.65 (.07)		
Reading Comprehension Task 1				.92 (.02)					.95 (.02)					.91 (.02)	
Reading Comprehension Task 2				.87 (.02)					.83 (.03)					.90 (.02)	
Narrative Writing					.94 (.02)					.84 (.03)					.90 (.03)
Argumentation Writing					.85 (.03)					.83 (.03)					.82 (.02)
Expository Writing					.76 (.04)					.84 (.03)					.79 (.05)
Inter-factor Correlations															
Morphological Awareness	.57	—				.58	—				.61	—			
Syntactic Processing	.61	.67	—			.47	.46	—			.51	.35	—		
Text Comprehension	.74	.84	.84	—		.72	.62	.68	—		.81	.76	.68	—	
Written Composition	.69	.78	.79	.85	—	.66	.62	.57	.77	—	.73	.65	.57	.85	—

Note. WM = working memory; MP = morphological awareness; SP = syntactic processing; TC = text comprehension; WC = written composition. All coefficients are significant at  $p < .001$ .

## Structural Equation Modeling

Structural equation models were used to examine hypothesized relations among the measured constructs (presented in Figure 1). We fit the model simultaneously to all three grades while constraining the factor loadings (except WNAR for Grade 5) and intercepts (except OSWM for Grade 5, and text comprehension task 2 for Grades 5 and 6) equal across grades as supported by the measurement invariance results. We first fit a model that specified morphological awareness, syntactic processing, and working memory as predictors of written composition. This model provided an excellent fit,  $\chi^2(df = 81) = 140.09, p < .001$ , CFI = .98, TLI = .97, RMSEA = .05 (90% CI: .04–.06), and SRMR = .04. We then added text comprehension as a mediating variable and reestimated the model. This model provided a satisfactory fit,  $\chi^2(df = 122) = 334.21, p < .001$ , CFI = .95, TLI = .93, RMSEA = .08 (90% CI: .07–.09), and SRMR = .05. The results from these two models are presented by grade in Table 6.

The first column presents bivariate latent variable correlations. Squaring these correlations gives an estimate of the shared variance between each predictor and written composition. The second column presents structure coefficients that were obtained when the latent variables were used as simultaneous predictors of written composition. The coefficients represent the independent contributions to prediction for each predictor. The third column presents structure coefficients for the predictors after text comprehension was added to the model as a potential mediator. The extent to which these structure coefficients were reduced compared to those without the mediator in the analysis indicates whether full, partial, or no mediation was occurring. The final two columns present estimate and bias corrected bootstrap 95% confidence interval of indirect effects of the predictors on written composition via the mediator variable of text comprehension. Significant indirect effects, noted by confidence intervals not containing zero, provide evidence of mediation.

The results of the first set of structural equation analyses indicated that morphological awareness, syntactic processing, and working memory were related to written composition individually and made independent contributions to prediction when considered as simultaneous predictors. When text comprehension was added as a potential mediator, the overall pattern of results was consistent with complete mediation. The predictive relations between the three predictors and written composition approached zero when text comprehension was added as a mediator. The bootstrap 95% confidence interval (CI) indicates that the mediation effect was significant for Grades 5 and 6, while marginally significant for Grade 4 (90% CI did not contain zero). The model accounted for approximately 75% of the variance in written composition.

Figures 2, 3, and 4 present the standardized path coefficients of the mediation model (as in Figure 1b) for the three grades. A chi-square difference test was conducted to examine whether each path was equal across grades (see Table 7). All paths were found equal except the path from working memory to text comprehension, which was found equal between Grades 5 and 6, but not between Grades 4 and 5 ( $p = .04$ ) nor between Grades 4 and 6 ( $p = .003$ ).

In summary, the results of structural equation models supported complete mediation of the effects of morphological awareness, syntactic processing, and working memory on written composition



Table 4  
Examination of Measurement Invariance Between Grades 4, 5, and 6

Model	Description	df	$\chi^2$	CFI	TLI	RMSEA	90% CI	SRMR	$\Delta df$	$\Delta \chi^2$
Model 1	Baseline Model	102	283.23***	.96	.93	.08	.07, .09	.04		
Model 2 (compared to Model 1)	Model with equal loadings	114	321.78***	.95	.93	.08	.07, .09	.06	12	37.90***
Model 3 (compared to Model 1)	Model with equal loadings except narrative writing of Grade 5	113	294.55***	.96	.93	.08	.07, .09	.05	11	14.03
Model 4 (compared to Model 3)	Model 3 + equal intercepts	125	329.59***	.95	.93	.08	.07, .09	.05	12	34.90***
Model 5 (compared to Model 3)	Model 3 + equal intercepts except operational span working memory of Grade 5 and text comprehension task 2 of Grades 5 and 6	122	299.95***	.96	.94	.07	.06, .09	.05	9	14.61

Note. CFI = comparative fit index; TLI = Tucker–Lewis coefficient; RMSEA = root-mean-square error of approximation; CI = confidence interval; SRMR = standardized root-mean-squared residual.  
\*\*\*  $p < .001$ .

via text processing. The results were largely consistent across grades.

Dominance Analysis

Table 8 presents the unique contribution in terms of proportion of variance explained by the four variables predicting written composition. The first column contains the total  $R^2$  for the corresponding subset model, and the remaining columns report the unique variance contribution added to that subset model. For example, for Grade 4, the subset model with WM demonstrates that 34% of written composition variance was accounted for by WM. After controlling for WM, the unique contribution to variance was 14% for MP and 8% for SP. In the subset model of WM-MP, the two predictors jointly accounted for 48% of variance, with 3% unique variance added by SP. Based on these unique contributions, we calculated the average contribution of a predictor as the mean of its average contribution over the subset models with the same number of predictors (Budescu, 1993). For all three grades, WM was the strongest predictor of written composition, uniquely contributing, on average, 20.67% of the variance for Grade 4, 17.00% of the variance for Grade 5, and 24.17% of the variance for Grade 6. This was followed by MP (Grade 4: 18.83%; Grade 5: 12.50%; Grade 6: 12.17%) and TC (Grade 4: 11.67%; Grade 5: 11.67%; Grade 6: 9.67%).

In Table 9, the first and second columns identify the two variables being compared; the third column is the value of dominance measure  $D_{ij}$  in the sample; the fourth column is the average value ( $\overline{D}_{ij}$ ) over the 1,000 bootstrap samples; and the fifth column is the standard error of the  $D_{ij}$  values. The next three columns

describe the distribution of  $D_{ij}$  over the 1,000 bootstrap samples, where  $P_{ij}$  is the proportion of samples in which  $X_i$  dominates  $X_j$ ,  $P_{ji}$  is the proportion of samples in which  $X_j$  dominates  $X_i$ , and  $P_{noij}$  is the proportion of samples in which the dominance could not be established. The last column is the reproducibility of the sample results, i.e., the proportion of bootstrap subsamples that agree with the tested sample results.

Examining the sample dominance values and reproducibility indices, working memory dominates morphological processing with high reproducibility for Grades 5 (82%) and 6 (98%), and with low reproducibility for Grade 4 (56%). Working memory dominates syntactic processing for all three grades with high reproducibility (92% for Grade 4, 83% for Grade 5, and 93% for Grade 6). The sample suggests that morphological processing dominated syntactic processing for Grade 4 (with 90% reproducibility), but undetermined for Grades 5 and 6.

In summary, the results of dominance analysis were that working memory dominated syntactic processing for all grades. For Grade 4, morphological awareness dominated syntactic processing. For Grades 5 and 6, working memory dominated morphological awareness. The other pairwise comparisons on unique contribution did not suggest reproducible dominance relationship.

Discussion

The goal of the study was to test opposing views about four issues concerning predictors of individual differences in Chinese written composition. We discuss results that address each issue before turning to limitations of our study and issues that are important to be addressed in future research.

Table 5  
Latent Means, Variances, and Latent Standardized Effect Sizes

Variable	Grade 4		Grade 5		Grade 6		Standardized effect size		
	<i>M</i>	Variance	<i>M</i>	Variance	<i>M</i>	Variance	Grade5–Grade4	Grade6–Grade5	Grade6–Grade4
Working Memory	0.00	32.33	2.43	20.48	4.06	21.63	0.47	0.36	0.78
Morphological Awareness	0.00	7.59	2.33	10.10	4.09	11.95	0.78	0.53	1.30
Syntactic Processing	0.00	5.37	1.99	7.12	3.87	3.94	0.80	0.80	1.80
Text Comprehension	0.00	37.05	4.67	51.33	6.75	51.05	0.70	0.29	1.01
Written Composition	0.00	194.70	9.74	229.76	19.41	176.18	0.67	0.68	1.43

Table 6

*Bivariate Correlations, Structure Coefficients, Structure Coefficients With Mediation, and Indirect Effects for Grades 4, 5, and 6*

Grade	Variable	Correlation	Structure coefficient	Structure coefficient with mediator	Indirect effect	
					Estimate	Bootstrap 95% CI
Grade 4	Working Memory	.69***	1.02*	0.31	0.57	(-0.25, 3.46)
	Morphological Awareness	.78***	1.96**	0.57	1.29	(-0.99, 5.77)
	Syntactic Processing	.79***	2.58*	0.79	1.59	(-0.17, 14.77)
Grade 5	Working Memory	.66***	1.91**	0.24	1.70	(0.82, 3.28)
	Morphological Awareness	.62***	1.48**	0.53	0.95	(0.34, 2.08)
	Syntactic Processing	.57***	1.73**	0.09	1.28	(0.56, 2.38)
Grade 6	Working Memory	.73***	1.62**	0.59	1.32	(0.06, 2.75)
	Morphological Awareness	.65***	1.51**	0.04	1.22	(0.04, 3.55)
	Syntactic Processing	.57***	2.01*	0.12	1.53	(0.22, 5.32)

Note. CI = confidence interval.

\*  $p < .01$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

1. *Distinct and measureable constructs in Chinese or just manifestations of general language ability?* Because previous studies typically represented morphological awareness, syntactic processing, and working memory as single indicator observed variables and did not include all three constructs, an important first step was to determine whether they represented distinct constructs or were merely manifestations of general language ability. We did this in the present study by including all three constructs and representing each as a latent variable with multiple indicators.

Based on the adequate model fits obtained for confirmatory factor analysis models that specified them as distinct yet potentially correlated abilities, and the fact that the obtained factor correlations ranged from .35 to .67, the results support morphological awareness, syntactic processing, and working memory as distinct yet correlated constructs, as opposed to just manifestations of general language ability. A single factor model specifying that the indicators of morphological awareness, syntactic processing,

and general language ability were indicators of general language ability resulted in substantially and significantly poorer model fits. The results then supported the view that morphological awareness, syntactic processing, and working memory are distinct and measurable constructs rather than just manifestations of general language ability. They are not independent, however, and their shared relations with general language ability are a likely source of their moderate intercorrelation.

2. *Are morphological awareness, syntactic processing, and working memory important predictors of Chinese written composition, and if so, what are the relative magnitudes and independence of their contributions to prediction?* Based on (a) their role in predicting English reading and to a lesser degree, English writing; (b) relations between reading and writing; and (c) characteristics of the Chinese writing system that place a premium on these three constructs, a theoretical rationale exists for expecting morphological awareness, syntactic processing, and working

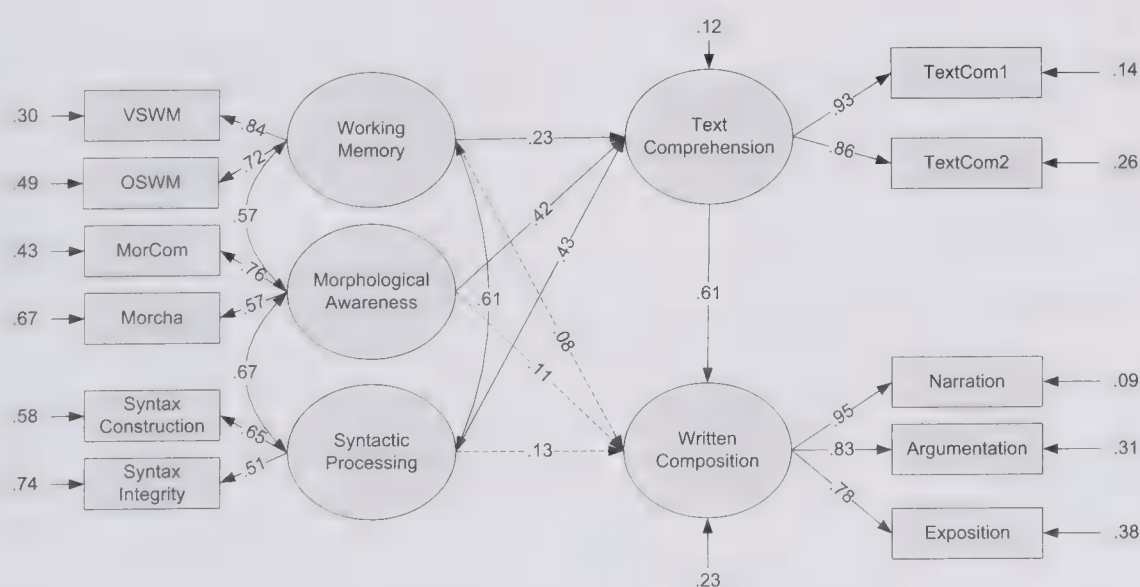


Figure 2. Structural equation model of Grade 4 showing standardized effects of working memory, morphological awareness, and syntactic processing on text comprehension (TextCom) and written composition. VSWM = verbal span working memory; OSWM = operation span working memory; MorCom = morphological compounding (from Leong & Ho, 2008); MorCha = morphological chain; Com = comprehension. All factor loadings, correlation coefficients, regression coefficients, and residual variances are significant at  $p < .03$ , except those indicated by dashed lines ( $ps > .37$ ).



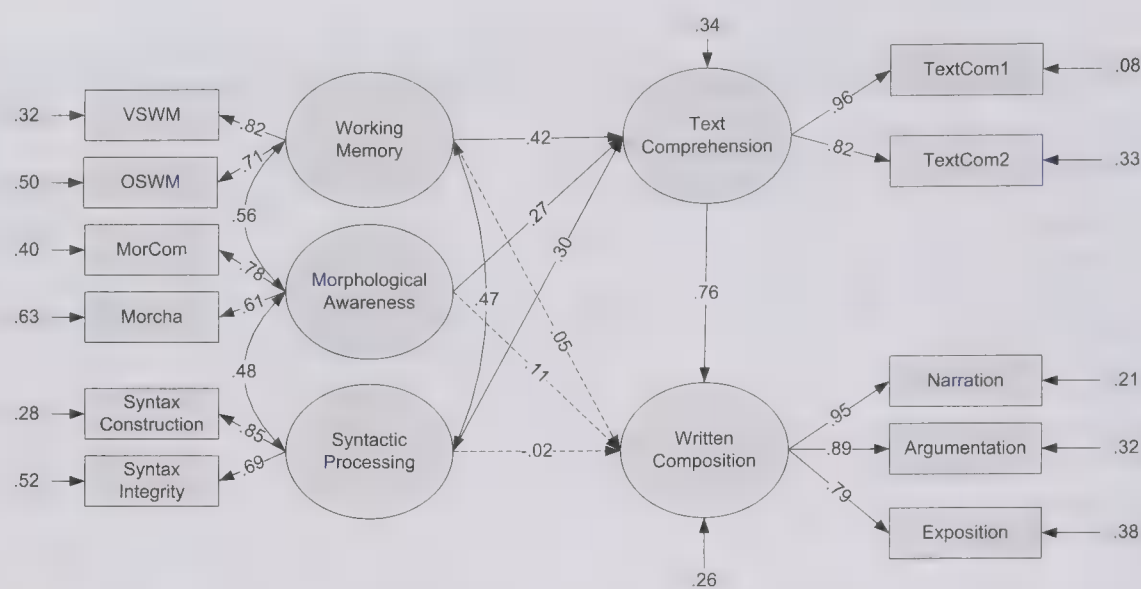


Figure 3. Structural equation model of Grade 5 showing standardized effects of working memory, morphological awareness, and syntactic processing on text comprehension (TextCom) and written composition. VSWM = verbal span working memory; OSWM = operation span working memory; MorCom = morphological compounding (from Leong & Ho, 2008); MorCha = morphological chain; Com = comprehension. All factor loadings, correlation coefficients, regression coefficients, and residual variances are significant at  $p < .01$ , except those indicated by dashed lines ( $ps > .15$ ).

memory to be important predictors of Chinese written composition. However, there is scant empirical evidence that tests this proposition. Results from the present study supported the importance of morphological awareness, syntactic processing, and working memory as important predictors of Chinese written composition. Factor correlations between the predictors of morphological awareness, syntactic processing, and working memory and the criterion of written composition obtained from the confirmatory

factor analyses, which are equivalent to bivariate regression coefficients, ranged from .6 to .8.

Finding morphological awareness to be an important predictor of Chinese written composition is consistent with previous studies that suggest it is related to learning to read in Chinese (Hao et al., 2013; Kuo & Anderson, 2006; Liu & McBride-Chang, 2010; Packard et al., 2006; Shu et al., 2006; Zhang et al., 2012) and predicts Chinese writing ability (Leong & Ho, 2008; Leong et al.,

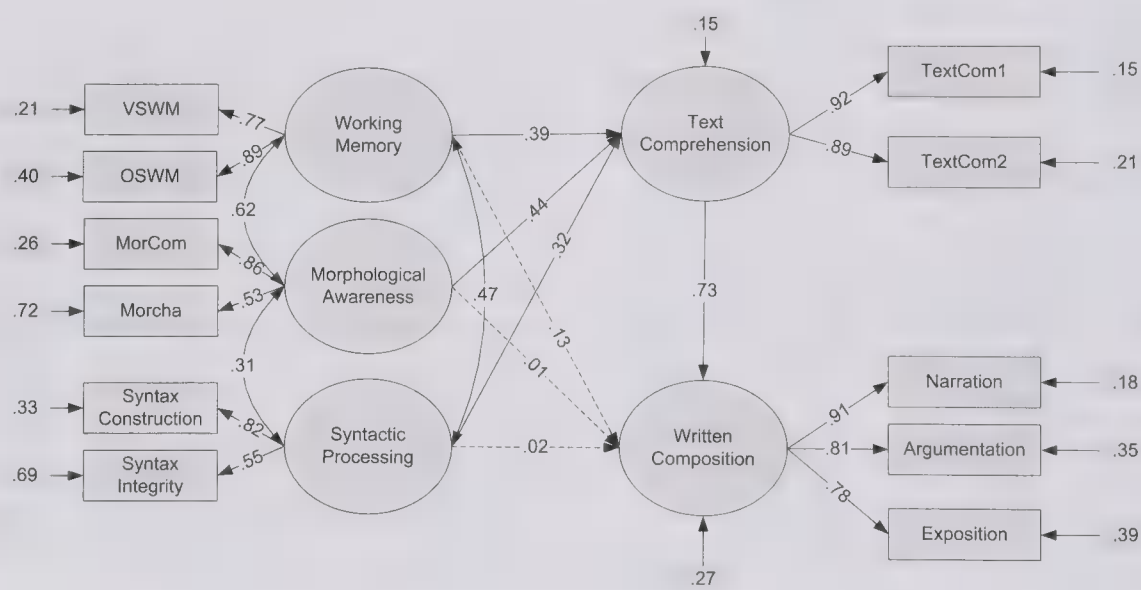


Figure 4. Structural equation model of Grade 6 showing standardized effects of working memory, morphological awareness, and syntactic processing on text comprehension (TextCom) and written composition. VSWM = verbal span working memory; OSWM = operation span working memory; MorCom = morphological compounding (from Leong & Ho, 2008); MorCha = morphological chain; Com = comprehension. All factor loadings, correlation coefficients, regression coefficients, and residual variances are significant at  $p < .002$ , except those indicated by dashed lines ( $ps > .28$ ).

Table 7  
Examination of Equality of Path Coefficients Between Grades 4, 5, and 6

Model	Description	df	$\chi^2$	$\Delta df$	$\Delta\chi^2$
Model 1	Baseline Model	122	334.22***		
Model 2 (compared to Model 1)	Model with equal WM→TC	124	342.11***	2	7.43*
Model 3 (compared to Model 1)	Model with equal MP→TC	124	334.13***	2	1.37
Model 4 (compared to Model 1)	Model with equal SP→TC	124	330.70***	2	0.85
Model 5 (compared to Model 1)	Model with equal TC→Writing	124	334.24***	2	0.60
Model 6 (compared to Model 1)	Model with equal WM→Writing	124	332.88***	2	0.45
Model 7 (compared to Model 1)	Model with equal MP→Writing	124	335.54***	2	1.32
Model 8 (compared to Model 1)	Model with equal SP→Writing	124	335.22***	2	0.99

Note. WM = working memory; TC = text comprehension; MP = morphological awareness; SP = syntactic processing.

\*  $p < .05$ . \*\*\*  $p < .001$ .

2013). Finding working memory to be an important predictor of Chinese written composition is consistent with results of previous research that has focused primarily on working memory in monolingual Chinese- and English-speaking children (Chung & McBride-Chang, 2011; Kellogg, 2001, 2004). Working memory may contribute to writing performance because of the need to hold information in short-term working memory while retrieving information from long-term memory (McCutchen, 2011; Vanderberg & Swanson, 2007). During this process, mental representation and focused manipulation of information are important. The role of syntactic processing appeared to be larger for higher relative to lower grades. This is consistent with the observation that more skilled writers apply their knowledge of syntax to their writing to a greater extent than do less skilled writers (Cromer & Wiener, 1966) and with the observation that knowledge of syntactic structures is necessary for processing higher level genres (Beers & Nagy, 2009, 2011).

Turning to relative magnitudes of prediction, the results of dominance analysis indicated that working memory was the strongest predictor of Chinese writing, followed by morphological awareness and syntactic processing, which were largely comparable with some trend for morphological awareness to dominate syntactic processing as a predictor. These results are comparable with research showing the importance of working memory as a

predictor of writing in English (Berninger, Abbott, et al., 2002; Fitzgerald & Shanahan, 2000; Graham, 2006; Shanahan, 2006).

Finally, we wanted to determine whether morphological awareness, syntactic processing, and working memory made independent contributions to prediction of Chinese written expression or whether their predictive relations were redundant, perhaps because they were correlated with language ability and language ability in turn predicted writing. The results of structural equation modeling supported the independence of their contribution to prediction. Significant structure coefficients were found for each predictor when they were included as simultaneous predictors of written composition (Table 6) without including text processing as a mediator.

3. *Are observed predictive relations mediated by text comprehension?* Given the similarities and differences between reading and writing discussed earlier, it was important to determine whether predictive relations between the three key constructs of morphological awareness, syntactic processing, and working memory and the dependent variable of Chinese written composition might be mediated by text comprehension. We therefore compared alternative models that proposed that predictive relations between morphological awareness, syntactic processing, and working memory were (a) unmediated, (b) partially mediated, or (c) fully mediated by text comprehension.

Table 8  
Dominance Analysis Results of Variables Predicting Writing

Subset model	Unique contribution of predictors to writing											
	Grade 4				Grade 5				Grade 6			
	$R^2$	WM	MP	SP	$R^2$	WM	MP	SP	$R^2$	WM	MP	SP
Models with one predictor												
WM	.34		.14	.08	.28		.08	.09	.39		.04	.04
MP	.32	.16		.08	.23	.13		.09	.24	.19		.10
SP	.24	.18	.17		.21	.15	.11		.18	.24	.15	
Models with two predictors												
WM-MP	.48			.03	.36			.05	.42			.04
WM-SP	.42		.09		.37		.05		.43		.03	
MP-SP	.40	.11			.32	.09			.34	.12		
Models with three predictors												
WM-MP-SP	.52				.42				.46			

Note. WM = working memory; MP = morphological awareness; SP = syntactic processing.



Table 9  
*The Sample Dominance and Their Means, Standard Errors, Probabilities, and Reproducibility Over 1,000 Bootstrap Samples*

Grade	<i>i</i>	<i>j</i>	<i>D<sub>ij</sub></i>	$\overline{D}_{ij}$	<i>SE(D<sub>ij</sub>)</i>	<i>P<sub>ij</sub></i>	<i>P<sub>ji</sub></i>	<i>P<sub>notij</sub></i>	Reproducibility
Grade 4	1	2	1.0	0.62	0.46	0.56	0.32	0.11	0.56
	1	3	1.0	0.95	0.19	0.92	0.02	0.06	0.92
	2	3	1.0	0.93	0.22	0.90	0.04	0.07	0.90
Grade 5	1	2	1.0	0.84	0.35	0.82	0.14	0.05	0.82
	1	3	1.0	0.85	0.34	0.83	0.13	0.04	0.83
	2	3	0.5	0.53	0.45	0.43	0.38	0.19	0.19
Grade 6	1	2	1.0	0.99	0.09	0.98	0.00	0.02	0.98
	1	3	1.0	0.96	0.15	0.93	0.01	0.06	0.93
	2	3	0.5	0.60	0.41	0.45	0.25	0.31	0.31

*Note.* 1 = Working Memory; 2 = Morphological Processing; 3 = Syntactic Processing.

Our results were consistent with the view that the predictive role of morphological awareness, syntactic processing, and working memory in accounting for individual differences in written composition is mediated through text comprehension. The mediation model accounted for approximately 75% of the variance in written composition. The results supported full rather than partial mediation and are consistent with other studies that suggest writing depends on reading (Ahmed et al., in press; Fitzgerald & Shanahan, 2000; Shanahan & Lomax, 1986). However, further research is necessary to support text comprehension as a true mediator. At a minimum, our results indicate that morphological awareness, syntactic processing, and working memory do not predict written composition independently of text comprehension. A true mediating role would require evidence that text comprehension actually facilitates written composition. Without further evidence from longitudinal and experimental studies, it is possible that the observed relation between text comprehension and written composition might be subserved by a third construct such as language or verbal aptitude.

It should also be noted that in the design of the study we asked our participants to write three different genres of composition—narration, argumentation and exposition—so as to provide as comprehensive a picture as possible of the students’ writing performance. Even though our intent was not to analyze the effects of our predictors on each kind of writing, we were also interested in the relative performance of the students. The results show the general trend of better performance of narratives, then expository writing followed by argumentation writing, grade for grade (Table 1). This differential performance by the Grades 4, 5, and 6 students is in keeping with the findings of the literature (Bereiter & Scardamalia, 1987; Langer, 1986). There is also evidence from recent reading psychology literature that different competencies contribute to children’s comprehension of narrative, expository and argumentation texts because of their different structure and different demands made on resource allocation (Best, Floyd, & McNamara, 2008; Reznitskaya et al., 2007). It is likely what is known for reading applies equally for writing (Englert, Stewart, & Hiebert, 1988).

For our specially designed text comprehension tasks with four narrative and 4 expository texts and the use of open-ended written comprehension tasks emphasizing inference, we also aimed at a broader portrayal of text comprehension. Our approach in designing the text comprehension tasks should address some of the

concerns raised about the influence of text and question types influencing reading comprehension (Eason, Goldberg, Young, Geist, & Cutting, 2012). What is not known is the mediating effect of particular genres of text on particular genres of writing. What is also not known is the effect of prior knowledge and knowledge utilization in writing. From inspection of the writing protocols and observation of the students it seemed that they were more intent on content generation and followed the task-execution model of knowledge telling, rather than the knowledge transformation of Bereiter and Scardamalia (1987).

The existing literature has not yet settled on a clear consensus about the nature and direction of relations between reading and writing (Aarnoutse, van Leeuwe & Verhoeven, 2005; Abbott et al., 2010; Babayiğit & Stainthorp, 2011; Berninger, Vaughan, et al., 2002; Caravolas, Hulme, & Snowling, 2001; Cataldo & Ellis, 1988; Lerkkanen, Rasku-Puttonen, Aunola, & Nurmi, 2004; Shanahan & Lomax, 1986; Sprenger-Charolles, Siegel, Bechenec, & Serniclaes, 2003). Text comprehension and written composition would seem to draw on similar linguistic and cognitive mechanisms and are likely to be mutually facilitative, but further studies are needed to understand their codevelopment.

4. *Developmental differences or invariance?* In the present study, we analyzed the data separately by grade to examine the extent to which our results varied by grade across the developmental range represented by fourth through sixth grades. The results supported developmental invariance on two levels. First, the measurement models were largely invariant across the three grades, supporting the assertion that the latent variables used to represent the constructs of interest were equivalent across the three grades. This enabled examination of changes in latent variable means across grades. Second, relations among the latent variables also were largely invariant across grades. These results indicate that the fourth through sixth grade students differed primarily in latent variable means, rather than in what the latent variables measured or how they were related with one another. These results are consistent with other recent studies that showed evidence of developmental invariance in writing (Guan et al., 2013; Wagner et al., 2011), although it is important to keep in mind the relatively limited developmental range represented by the fourth through sixth grades.

## Limitations, Implications, and Future Directions

One limitation of our study is relying on a cross-sectional rather than a longitudinal design. Although this design has the virtue of a relatively larger number of participants and a shorter duration compared to a longitudinal design, examination of developmental differences is confounded with potential cohort effects. A longitudinal design would be particularly helpful for a more rigorous test of mediational relations (Abbott, Amtmann, & Munson, 2006). A second limitation is the limited developmental range represented by including participants from grades four through six. Although writing performance does change over this period of time, a larger developmental range would be helpful in studying what changes and what does not with development. A third limitation is the limited nature of our writing tasks. The methods for scoring quality of writing and comprehension were not typical and that for the quality scores some of the reliabilities were less than might be desired. Also, we did not incorporate important topics such as the processes involved in planning, formulating ideas, editing, and revising them to form coherent and cohesive written texts; writing for different purposes; and discourse knowledge about forms of writing (Graham, 2006; Graham & Harris, 2007; Graham & Perin, 2007; Olinghouse & Graham, 2009). Our writing tasks were group administered, which makes it possible that the writing behavior of a given student was influenced by the surrounding context of other students. Further, we did not control for the effect due to legibility before we scored the students' writing task (Graham, Harris, & Hebert, 2011). A meta-analysis by these authors suggests that legibility has a large effect on scoring quality of writing (Graham, & Hebert, 2011). Another limitation is that all students were from a single school, although this might assure participants coming from similar socioeconomic background and language/literacy experience. We also acknowledge that our results apply to normally developing writers and may not apply to students with impairments in writing or other aspects of language.

Despite its limitations, the current study provides greater contributions and practical implications to the field of educational psychology, educational practice, and possibly educational policy. First, the study might be one of the first that established the important measureable predictors for Chinese written composition. Second, by conducting dominance analysis, the study might be one of the first that revealed the relative magnitudes and independent contributions of each unique linguistic and cognitive factor to written composition. In writing practices, the teachers will be informed of how to focus on their elements of writing instructions to improve students' writing performance. The third contribution is to theories of educational psychology of writing research. The study conducted confirmatory factor analysis to distinguish our three-factor model of writing (morphological awareness, syntactic processing and working memory) from the single factor model of general language ability. Fourth, we compared theoretically the alternative models of unmediation, partial mediation, and full mediation of text comprehension of these three key constructs and Chinese written composition. As well, we provided theoretically based empirical evidence to show that the predictive relations between the three key constructs of linguistic and cognitive factors and the dependent variable of Chinese written composition might be mediated by text comprehension. This provides a potential alternative view of how we could address the predictive relations

among reading and writing variables. Our fifth contribution is to educational policy. This relates to ways of assessing reading and writing, and the predictive relations between linguistic and cognitive measures mediated by text comprehension. The results of our study might suggest a blueprint of reading and writing for educational policy makers.

In future studies, it will be important to consider different genres of written composition more specifically, as they may make different demands on planning, translating and review processes and on cognitive resources such as working memory that underlie them (Kellogg, 2001; Torrance & Jeffery, 1999). Finally, there is a need for randomized controlled trials of instructional approaches and interventions directed towards improving writing skill (Cutler & Graham, 2008).

## References

- Aarnoutse, C., van Leeuwe, J., & Verhoeven, L. (2005). Early literacy from a longitudinal perspective. *Educational Research and Evolution, 11*, 253–275. doi:10.1080/08993400500101054
- Abbott, R. D., Amtmann, D., & Munson, J. (2006). Statistical analysis for field experiments and longitudinal data in writing research. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 374–385). New York, NY: Guilford Press.
- Abbott, R. D., & Berninger, V. W. (1993). Structural equation modeling of relationships among developmental skills and writing skills in primary- and intermediate-grade writers. *Journal of Educational Psychology, 85*, 478–508. doi:10.1037/0022-0663.85.3.478
- Abbott, R. D., Berninger, V. W., & Fayol, M. (2010). Longitudinal relationships of levels of language in writing and between writing and reading in Grades 1 to 7. *Journal of Educational Psychology, 102*, 281–298. doi:10.1037/a0019318
- Ahmed, Y., Wagner, R. K., & Lopez, D. (in press). Developmental relations between reading and writing at the word, sentence, and text levels: A latent change score analysis. *Journal of Educational Psychology*.
- Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods, 8*, 129–148. doi:10.1037/1082-989X.8.2.129
- Babayigit, S., & Stainthorp, R. (2011). Modeling the relationships between cognitive-linguistic skills and literacy skills: New insights from a transparent orthography. *Journal of Educational Psychology, 103*, 169–189. doi:10.1037/a0021671
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182. doi:10.1037/0022-3514.51.6.1173
- Bazerman, C. (2008). *Handbook of research on writing: History, society, school, individual, text*. New York, NY: Erlbaum.
- Beers, S. F., & Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality measures? Which genre? *Reading and Writing: An Interdisciplinary Journal, 22*, 185–200. doi:10.1007/s11145-007-9107-5
- Beers, S. F., & Nagy, W. E. (2011). Writing development in four genres from grade three to seven: Syntactic complexity and genre differentiation. *Reading and Writing: An Interdisciplinary Journal, 24*, 183–202. doi:10.1007/s11145-010-9264-9
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246. doi:10.1037/0033-2909.107.2.238
- Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.
- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606. doi:10.1037/0033-2909.88.3.588



- Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Erlbaum.
- Berman, R. A., & Nir-Sagiv, B. (2007). Comparing narrative and expository text construction across adolescence: A developmental paradox. *Discourse Processes*, 43, 79–120.
- Berninger, V. W., Abbott, R. D., Abbott, S. P., Graham, S., & Richards, T. (2002). Writing and reading: Connections between language by hand and language by eye. *Journal of Learning Disabilities*, 35, 39–56. doi:10.1177/002221940203500104
- Berninger, V. W., & Chanquoy, L. (2012). What writing is and how it changes across early and middle childhood development: A multidisciplinary perspective. In E. Grigorenko, E. Mambrino, & D. Preiss (Eds.), *Writing: A mosaic of new perspectives* (pp. 65–84). New York, NY: Psychology Press.
- Berninger, V. W., Vaughan, K., Graham, S., Abbott, R. D., Begay, K., Coleman, K. B., . . . Hawkins, J. M. (2002). Teaching spelling and composition alone and together: Implications for the simple view of writing. *Journal of Educational Psychology*, 94, 291–304. doi:10.1037/0022-0663.94.2.291
- Berninger, V. W., & Winn, W. D. (2006). Implications of advancements in brain research and technology for writing development, writing instruction, and educational evolution. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 96–114). New York, NY: Guilford Press.
- Best, R. M., Floyd, R. G., & McNamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology*, 29, 137–164. doi:10.1080/02702710801963951
- Bi, Y., Han, Z., & Zhang, Y. (2009). Reading does not depend on writing, even in Chinese. *Neuropsychologia*, 47, 1193–1199. doi:10.1016/j.neuropsychologia.2008.11.006
- Britton, B. K. (1994). Understanding expository text: Building mental structure to induce insights. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 641–674). New York, NY: Academic Press.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.
- Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, 114, 542–551. doi:10.1037/0033-2909.114.3.542
- Caravolas, M., Hulme, C., & Snowling, M. (2001). The foundations of spelling ability: Evidence from a 3-year longitudinal study. *Journal of Memory and Language*, 45, 751–774. doi:10.1006/jmla.2000.2785
- Cataldo, S., & Ellis, N. (1988). Interactions in the development of spelling, reading and phonological skills. *Journal of Research in Reading*, 11, 86–109. doi:10.1111/j.1467-9817.1988.tb00153.x
- Chik, P. P.-m., Ho, C. S.-h., Yeung, P.-s., Chan, D. W.-o., Chung, K. K.-h., Luan, H., . . . Lau, W. S.-y. (2012). Syntactic skills in sentence reading comprehension among Chinese elementary school children. *Reading and Writing: An Interdisciplinary Journal*, 25, 679–699. doi:10.1007/s11145-010-9293-4
- Choi, J., Fan, W., & Hancock, G. R. (2009). A note on confidence intervals for two group latent mean effect size measures. *Multivariate Behavioral Research*, 44, 396–406. doi:10.1080/00273170902938902
- Chung, K. K. H., & McBride-Chang, C. (2011). Executive functioning skills uniquely predict Chinese word reading. *Journal of Educational Psychology*, 103, 909–921. doi:10.1037/a0024744
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Courville, T., & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles:  $\beta$  is not enough. *Educational and Psychological Measurement*, 61, 229–248.
- Cromer, W., & Wiener, M. (1966). Idiosyncratic response patterns among good and poor readers. *Journal of Consulting Psychology*, 30, 1–10. doi:10.1037/h0022885
- Cutler, L., & Graham, S. (2008). Primary grade writing instruction: A national survey. *Journal of Educational Psychology*, 100, 907–919. doi:10.1037/a0012656
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior*, 19, 450–466. doi:10.1016/S0022-5371(80)90312-6
- Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, 3, 422–433. doi:10.3758/BF03214546
- Eason, S. H., Goldberg, L. F., Young, C. M., Geist, M. C., & Cutting, L. E. (2012). Reader-text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology*, 104, 515–528. doi:10.1037/a0027182
- Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, 128, 309–331. doi:10.1037/0096-3445.128.3.309
- Englert, C. S., Stewart, S. R., & Hiebert, E. H. (1988). Young writers' use of text structure in expository text generation. *Journal of Educational Psychology*, 80, 143–151. doi:10.1037/0022-0663.80.2.143
- Fabb, N. (1998). Compounding. In A. Spencer & M. Zwicky (Eds.), *The handbook of morphology* (pp. 66–83). Oxford, England: Blackwell.
- Fitzgerald, J., & Shanahan, T. (2000). Reading and writing relations and their development. *Educational Psychologist*, 35, 39–50. doi:10.1207/S15326985EP3501\_5
- Foorman, B. R., Arndt, E. J., & Crawford, E. C. (2011). Important constructs in literacy learning across disciplines. *Topics in Language Disorders*, 31, 73–83. doi:10.1097/TLD.0b013e31820a0b86
- Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *RASE: Remedial & Special Education*, 7, 6–10. doi:10.1177/074193258600700104
- Graham, S. (2006). Writing. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 457–478). Mahwah, NJ: Erlbaum.
- Graham, S., & Harris, K. R. (Eds.). (2000). Writing development: The role of cognitive, motivational, and social/contextual factors [Special issue]. *Educational Psychologist*, 35(1).
- Graham, S., & Harris, K. R. (2007). Best practices in teaching planning. In S. Graham, C. A. MacArthur, & J. Fitzgerald (Eds.), *Best practices in writing instruction* (pp. 119–140). New York, NY: Guilford Press.
- Graham, S., & Harris, K. R. (2009). Almost 30 years of writing research: Making sense of it all with *The Wrath of Khan*. *Learning Disabilities Research & Practice*, 24, 58–68. doi:10.1111/j.1540-5826.2009.01277.x
- Graham, S., Harris, K. R., & Hebert, M. (2011). It is more than just the message: Analysis of presentation effects in scoring writing. *Focus on Exceptional Children*, 44, 1–12.
- Graham, S., & Hebert, M. (2011). Writing to read: A meta-analysis of the impact of writing and writing instruction on reading. *Harvard Educational Review*, 81, 710–744.
- Graham, S., & Perin, D. (2007). What we know, what we still need to know: Teaching adolescents to write. *Scientific Studies of Reading*, 11, 313–335. doi:10.1080/10888430701530664
- Grigorenko, E., Mambrino, E., & Preiss, D. (Eds.). (2012). *Writing: A mosaic of new perspectives*. New York, NY: Psychology Press.
- Guan, C. Q., Liu, Y., Ye, F., Chan, D. H. L., & Perfetti, C. A. (2011). Writing strengthens orthography and alphabetic-coding strengthens phonology in learning to read Chinese. *Journal of Educational Psychology*, 103, 509–522. doi:10.1037/a0023730

- Guan, C. Q., Ye, F., Meng, W., & Leong, C. K. (2013). Are poor Chinese text comprehenders also poor in written composition? *Annals of Dyslexia*, 63, 217–238. doi:10.1007/s11881-013-0081-0
- Guan, C. Q., Ye, F., Wagner, R. K., & Meng, W. (2013). Developmental and individual differences in Chinese writing. *Reading and Writing: An Interdisciplinary Journal*, 26, 1031–1056. doi:10.1007/s11145-012-9405-4
- Hao, M., Chen, X., Dronjic, V., Shu, H., & Anderson, R. C. (2013). The development of young Chinese children's morphological awareness: The role of semantic relatedness and morpheme type. *Applied Psycholinguistics*, 34, 45–67. doi:10.1017/S0142716411000609
- Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1–27). Mahwah, NJ: Erlbaum.
- Hayes, J. R. (2006). New directions in writing theory. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 28–40). New York, NY: Guilford Press.
- Hayes, J. R., & Chenoweth, N. A. (2006). Is working memory involved in the transcribing and editing of texts? *Written Communication*, 23, 135–149. doi:10.1177/0741088306286283
- Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Hillsdale, NJ: Erlbaum.
- Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal*, 2, 127–160. doi:10.1007/BF00401799
- Hoskyn, M., & Swanson, H. L. (2003). The relationship between working memory and writing in younger and older adults. *Reading and Writing: An Interdisciplinary Journal*, 16, 759–784. doi:10.1023/A:1027320226283
- Hu, L.-T., & Bentler, P. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). London, England: Sage.
- Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55. doi:10.1080/10705519909540118
- Jenkins, J. R., Johnson, E., & Hileman, J. (2004). When is reading also writing: Sources of individual differences on the new reading performance assessment. *Scientific Studies of Reading*, 8, 125–151. doi:10.1207/s1532799xssr0802\_2
- Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grade. *Journal of Educational Psychology*, 80, 437–447. doi:10.1037/0022-0663.80.4.437
- Juel, C., Griffith, P., & Gough, P. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. *Journal of Educational Psychology*, 78, 243–255. doi:10.1037/0022-0663.78.4.243
- Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences and applications* (pp. 57–71). Mahwah, NJ: Erlbaum.
- Kellogg, R. T. (1999). Components of working memory in text production. In M. Torrance & G. C. Jeffery (Eds.), *The cognitive demands of writing: Processing capacity and working memory in text production* (pp. 43–61). Amsterdam, the Netherlands: Amsterdam University Press.
- Kellogg, R. T. (2001). Competition of working memory among writing processes. *The American Journal of Psychology*, 114, 175–191. doi:10.2307/1423513
- Kellogg, R. T. (2004). Working memory components in written sentence generation. *The American Journal of Psychology*, 117, 341–361. doi:10.2307/4149005
- Kellogg, R. T., Whiteford, A. P., Turner, C. E., Cahill, M., & Mertens, A. (2013). Working memory in written composition: An evaluation of the 1996 model. *Journal of Writing Research*, 5, 159–190.
- Kintsch, W., & Kintsch, E. (2005). Comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 71–92). Mahwah, NJ: Erlbaum.
- Kirby, J. R., Deacon, S. H., Bowers, P. N., Izenberg, L., Wade-Wolley, L., & Parrila, R. (2012). Children's morphological awareness and reading. *Reading and Writing: An Interdisciplinary Journal*, 25, 389–410. doi:10.1007/s11145-010-9276-5
- Kuo, L.-j., & Anderson, R. C. (2006). Morphological awareness and learning to read: A cross-language perspective. *Educational Psychologist*, 41, 161–180. doi:10.1207/s15326985ep4103\_3
- Langer, J. A. (1986). *Children reading and writing: Structures and strategies*. Norwood, NJ: Ablex.
- Leong, C. K., & Ho, M. K. (2008). The role of lexical knowledge and related linguistic components in typical and poor language comprehenders of Chinese. *Reading and Writing: An Interdisciplinary Journal*, 21, 559–586. doi:10.1007/s11145-007-9113-7
- Leong, C. K., Ho, M. K., Chang, J., & Hau, K. T. (2013). Differential importance of language components in determining secondary school students' Chinese reading literacy performance. *Language Testing*, 30, 419–439. doi:10.1177/0265532212469178
- Leong, C. K., Tse, S. K., Loh, K. Y., & Hau, K. T. (2008). Text comprehension in Chinese children: Relative contribution of verbal working memory, pseudoword reading, rapid automatized naming, and onset-rime phonological segmentation. *Journal of Educational Psychology*, 100, 135–149. doi:10.1037/0022-0663.100.1.135
- Lerkanen, M. K., Rasku-Puttonen, H., Aunola, K., & Nurmi, J. E. (2004). Developmental dynamics of phonemic awareness and reading performance during the first year of primary school. *Journal of Early Childhood Research*, 2, 139–156. doi:10.1177/1476718X04042972
- Li, C. N., & Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar*. Berkeley, CA: University of California Press.
- Liu, P. D., & McBride-Chang, C. (2010). What is morphological awareness? Tapping lexical compounding awareness in Chinese third graders. *Journal of Educational Psychology*, 102, 62–73. doi:10.1037/a0016933
- MacArthur, C. A., Graham, S., & Fitzgerald, J. (Eds.). (2006). *Handbook of writing research*. New York, NY: Guilford Press.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1, 130–149. doi:10.1037/1082-989X.1.2.130
- McCutchen, D. (2000). Knowledge acquisition, processing efficiency, and working memory: Implications for a theory of writing. *Educational Psychologist*, 35, 13–23. doi:10.1207/S15326985EP3501\_3
- McCutchen, D. (2006). Cognitive factors in the development of children's writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 115–130). New York, NY: Guilford Press.
- McCutchen, D. (2011). From novice to expert: Implications of language skills and writing-relevant knowledge for memory during the development of writing skill. *Journal of Writing Research*, 3, 51–68.
- Miller, J., & Chapman, R. (2001). *Systematic analysis of language transcripts* (Version 7.0) [Computer software]. Madison, WI: Waisman Center, University of Wisconsin—Madison.
- Muthén, L. K., & Muthén, B. O. (Eds.). (2010). *Mplus: Statistical analysis with latent variables* (Version 6.1). Los Angeles, CA: Muthén & Muthén.
- NIES. (2012). *Initial success in assessment and intervention on LD at GaoQiao Primary School in NingBo*. Retrieved from [http://www.nies.net.cn/ky/syq/nbyz/jgcz/201211/t20121101\\_306943.html](http://www.nies.net.cn/ky/syq/nbyz/jgcz/201211/t20121101_306943.html)
- Olinghouse, N. G., & Graham, S. (2009). The relationship between the discourse knowledge and the writing performance of elementary-grade



- students. *Journal of Educational Psychology*, 101, 37–50. doi:10.1037/a0013462
- Packard, J. L., Chen, X., Li, W., Wu, X., Gaffney, J. S., Li, H., & Anderson, R. C. (2006). Explicit instruction in orthographic structure and word morphology helps Chinese children learn to write characters. *Reading and Writing: An Interdisciplinary Journal*, 19, 457–487. doi:10.1007/s11145-006-9003-4
- Perfetti, C. A., & Guan, C. Q. (2012, April). *Effect of repeated writing practice*. In C. Q. Guan (Chair), *Written language studies across culture*. Symposium conducted at the American Educational Research Association annual meeting, Vancouver, Canada.
- Ployhart, R. E., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods*, 7, 27–65. doi:10.1177/1094428103259554
- Reznitskaya, A., Anderson, R. C., & Kuo, L.-J. (2007). Teaching and learning argumentation. *The Elementary School Journal*, 107, 449–472. doi:10.1086/518623
- Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis: A Festschrift for Heinz Neudecker* (pp. 233–247). London, England: Kluwer Academic. doi:10.1007/978-1-4615-4603-0\_17
- Satorra, A., & Bentler, P. M. (2001). A scaled differences chi-square test statistic for moment structure analysis. *Psychometrika*, 66, 507–514. doi:10.1007/BF02296192
- Schmandt-Besserat, D., & Erard, M. (2008). Origins and forms of writing. In C. Bazerman (Ed.), *Handbook of research on writing* (pp. 7–22). New York, NY: Erlbaum.
- Shanahan, T. (1984). Nature of the reading–writing relation: An exploratory multivariate analysis. *Journal of Educational Psychology*, 76, 466–477. doi:10.1037/0022-0663.76.3.466
- Shanahan, T. (2006). Relations among oral language, reading, and writing development. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 171–183). New York, NY: Guilford Press.
- Shanahan, T., & Lomax, R. C. (1986). An analysis and comparison of theoretical models of reading–writing relationship. *Journal of Educational Psychology*, 78, 116–123. doi:10.1037/0022-0663.78.2.116
- Shu, H., McBride-Chang, C., Wu, S., & Liu, H. (2006). Understanding Chinese developmental dyslexia: Morphological awareness as a core cognitive construct. *Journal of Educational Psychology*, 98, 122–133. doi:10.1037/0022-0663.98.1.122
- Sprenger-Charolles, L., Siegel, L. S., Bechenec, D., & Serniclaes, W. (2003). Development of phonological and orthographic processing in reading aloud, in silent reading, and in spelling: A four-year longitudinal study. *Journal of Experimental Child Psychology*, 84, 194–217. doi:10.1016/S0022-0965(03)00024-9
- Swanson, H. L. (1992). Generality and modifiability of working memory among skilled and less skilled readers. *Journal of Educational Psychology*, 84, 473–488. doi:10.1037/0022-0663.84.4.473
- Swanson, H. L., & Berninger, V. W. (1996). Individual differences in children's writing: A function of working memory or reading or both processes? *Reading and Writing: An Interdisciplinary Journal*, 8, 357–383. doi:10.1007/BF00395114
- Tan, L. H., Spinks, J. A., Eden, G. F., Perfetti, C. A., & Siok, W. T. (2005). Reading depends on writing, in Chinese. *PNAS Proceedings of the National Academy of Science of the United States of America*, 102, 8781–8785. doi:10.1073/pnas.0503523102
- Tierney, R. J., & Shanahan, T. (1991). Research on the reading–writing relationship: Interactions, transactions, and outcomes. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 246–280). Mahwah, NJ: Erlbaum.
- Torrance, M., & Galbraith, D. (2006). The processing demands of writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 67–80). New York, NY: Guilford Press.
- Torrance, M., & Jeffery, G. C. (Eds.). (1999). *The cognitive demands of writing: Processing capacity and working memory in text production*. Amsterdam, the Netherlands: Amsterdam University Press. doi:10.5117/9789053563083
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4–70. doi:10.1177/109442810031002
- Vandenberg, R., & Swanson, H. L. (2007). Which components of working memory are important in the writing process? *Reading and Writing: An Interdisciplinary Journal*, 20, 721–752. doi:10.1007/s11145-006-9046-6
- Wagner, R. K., Puranik, C. S., Foorman, B., Foster, E., Wilson, L. G., Tschinkel, E., & Kantor, P. T. (2011). Modeling the development of written language. *Reading and Writing: An Interdisciplinary Journal*, 24, 203–220. doi:10.1007/s11145-010-9266-7
- Yan, C. M. W., McBride-Chang, C., Wagner, R. K., Zhang, J., Wong, A. M. Y., & Shu, H. (2012). Writing quality in Chinese children: Speed and fluency matter. *Reading and Writing: An Interdisciplinary Journal*, 25, 1499–1521. doi:10.1007/s11145-011-9330-y
- Yeung, P.-S., Ho, C. S.-H., Chik, P. P.-M., Lo, L.-Y., Luan, H., Chan, D. W.-O., & Chung, K. K.-H. (2011). Reading and spelling Chinese among beginning readers: What skills make a difference? *Scientific Studies of Reading*, 15, 285–313. doi:10.1080/10888438.2010.482149
- Zhang, J., Anderson, R. C., Wang, Q., Packard, J., Wu, X., Tang, S., & Ke, X. (2012). Insight into the structure of compound words among speakers of Chinese and English. *Applied Psycholinguistics*, 33, 753–779. doi:10.1017/S0142716411000555

Received March 10, 2012

Revision received January 13, 2014

Accepted January 15, 2014 ■

# Impact of a Teacher-Led Intervention on Preference for Self-Regulated Learning, Finding Main Ideas in Expository Texts, and Reading Comprehension

Heidrun Stoeger and Christine Sontag  
University of Regensburg

Albert Ziegler  
University of Erlangen–Nuremberg

We examined the impact of a teacher-led intervention, implemented during regular classroom instruction and homework, on fourth-grade students' preference for self-regulated learning, finding main ideas in expository texts, and reading comprehension. In our quasi-experimental study with intact classrooms, (a) students ( $n = 266$ , 12 classrooms) who received regular classroom instruction (REG) were compared with (b) students ( $n = 268$ , 12 classrooms) who were taught text reduction strategies (TEXT) and (c) students ( $n = 229$ , 9 classrooms) who were introduced to text reduction strategies within the framework of a 7-step cyclical model of self-regulated learning (SRL + TEXT). Participating classrooms were semi-randomly assigned to 1 of the 3 conditions, with the restriction that teachers from one school could not be in different intervention conditions. Both in their posttest and follow-up test results (11 weeks after the intervention), SRL + TEXT students showed a stronger preference for self-regulated learning than students of the 2 other groups. The SRL + TEXT students also identified more main ideas over the course of the intervention. Positive effects on reading comprehension in a standardized test were restricted to students without migration background.

**Keywords:** self-regulated learning, strategy instruction, text reduction strategies, reading comprehension, intervention study

**Supplemental materials:** <http://dx.doi.org/10.1037/a0036035.supp>

Self-regulated learning represents a key skill in our rapidly changing society and one that needs to be taught and practiced as early as possible (Council of the European Union, 2002). The substantial number of effective interventions focusing on self-regulated learning for elementary-school students supports this fact. However, teacher-led interventions produce effect sizes smaller than those of researcher-led interventions (Dignath & Büttner, 2008). Yet teacher-led interventions are particularly important in that they are well suited for encouraging knowledge transfer, as the transfer of self-regulated learning skills from the context in which they were acquired to other domains and contexts works best when the skills are introduced and taught in multiple authentic learning settings (Dignath & Büttner, 2008; Hattie, Biggs, & Purdie, 1996). We, therefore, designed a teacher-led

intervention for self-regulated learning for elementary school students that (a) is appropriate for regular classroom instruction and for homework (the two most important scholastic learning settings) and (b) can be applied in multiple subjects.

## Theoretical and Empirical Background

Meta-analyses indicate that for elementary-school settings, self-regulation interventions based on social cognitive theory are among the most effective (Dignath & Büttner, 2008; Dignath, Buettner, & Langfeldt, 2008). In his frequently cited, social-cognitive-theory-based model, Zimmerman (1989, 2000) divides the self-regulation process into three subsequent phases: a *forethought phase*, a *performance or volitional-control phase*, and a *self-reflection phase*.

The forethought phase encompasses those prerequisite processes that precede actions and learning efforts. The performance or volitional-control phase includes processes that are important during learning and influence one's focus and behavior. During the self-reflection phase, which begins after learning and concludes the cyclical model by Zimmerman (2000), learners evaluate the outcome of their learning. Processes occurring during the self-reflection phase influence the next forethought phase. Each phase within the model brings together numerous cognitive, metacognitive, and motivational aspects (for an overview, cf. Zimmerman, 2000).

By conceptualizing optimal self-regulated learning, models such as Zimmerman's (2000) provide a basis for designing interven-

---

This article was published Online First March 17, 2014.

Heidrun Stoeger and Christine Sontag, Chair for School Research, School Development, and Evaluation, University of Regensburg; Albert Ziegler, Chair for Educational Psychology, University of Erlangen–Nuremberg.

We would like to thank Teresa Greindl for her assistance in data collection and data entry, and Daniel Patrick Balestrini and Sebastian Suggate for language proofreading.

Correspondence concerning this article should be addressed to Heidrun Stoeger, Chair for School Research, School Development, and Evaluation, University of Regensburg, 93040 Regensburg, Germany. E-mail: [heidrun.stoeger@ur.de](mailto:heidrun.stoeger@ur.de)



tions and for conducting research on self-regulated learning. Research findings indicate that acquainting intervention participants with an intervention's theoretical model improves both its effectiveness and transfer (Salomon & Perkins, 1989; Stahl, Simpson, & Hayes, 1992). However, as theoretical models such as that of Zimmerman (2000) include numerous aspects in each phase and are therefore relatively complex, a simplified version should be taught to intervention participants (Stoeger & Ziegler, 2008a, 2011; Zimmerman, Bonner, & Kovach, 1996). Model simplification is, moreover, particularly important when interventions target children in elementary school (Zimmerman, 1990).

For our intervention, we chose a simplified seven-step cyclical normative model of self-regulated learning (Ziegler & Stoeger, 2005) that reflects a limited number of important aspects from the Zimmerman model (cf. online supplemental material, Figure S1). The simplified model stresses those cognitive and metacognitive aspects for which there are promising results from earlier interventions with elementary school students (Dignath & Büttner, 2008; Stoeger & Ziegler, 2008a). This model places less emphasis on motivational aspects, as motivation issues appear to play a greater role in interventions for secondary school students (Dignath & Büttner, 2008). The first three steps of the intervention model represent aspects contained within Zimmerman's (2000) forethought phase. They are *self-assessment*, *goal setting*, and *strategic planning*. The next three steps—*strategy implementation*, *strategy monitoring*, and *strategy adjustment*—represent aspects contained within Zimmerman's (2000) performance or volitional-control phase. These three steps constitute an internal cycle within the larger seven-step cyclical model and can be applied to various cognitive strategies (e.g., organizational strategies, rehearsal strategies; cf. Weinstein & Mayer, 1986). The final step in the seven-step cycle of self-regulated learning, *outcome evaluation*, is derived from the third phase of Zimmerman's (2000) model. As in Zimmerman's (2000) model, this final step influences the way students work through the cycle of self-regulated learning the next time.

In addition to the choice of the underlying theoretical model, several other features of self-regulated learning interventions have been associated with producing particularly large effect sizes. Effect sizes are larger if interventions emphasize the benefit of strategy use and provide systematic feedback (Dignath & Büttner, 2008; Hattie & Timperley, 2007; Schunk & Rice, 1987). Furthermore, evidence indicates that introducing self-regulated learning with concrete subject matter in real-life settings is particularly effective and helpful for improving transfer to other tasks and situations (Dignath & Büttner, 2008; Hattie et al., 1996). Additionally, interventions are especially effective when they simultaneously address both in-class instruction and homework contexts (Ramdass & Zimmerman, 2011; Stoeger & Ziegler, 2011). Finally, the duration of an intervention has an influence on its overall efficacy and the extent to which learners succeed in transferring a given skill from the context in which it was taught into new subject areas and learning contexts (Alexander, Graham, & Harris, 1998; Pressley, Graham, & Harris, 2006).

### Present Research

We sought to build upon current research findings by developing a 7-week teacher-led training program that students apply

during regular classroom instruction and homework. We developed the intervention for fourth grade in accordance with Bavarian state curriculum guidelines that explicitly mandate the introduction of self-regulation skills during fourth grade (Bayerisches Staatsministerium für Unterricht und Kultus, 2000). Basic science<sup>1</sup> and reading instruction were chosen as content areas. Based on the research reported earlier, we make the assumption that introducing self-regulated learning in the context of two school subjects and during in-class instruction and homework should facilitate the transfer of self-regulated learning skills.

In accordance with this content focus, we selected text reduction strategies for Steps 4 through 6 of the seven-step cycle of self-regulated learning. The training program introduces students to three reduction strategies that are useful for identifying and displaying main ideas: *underlining and copying main ideas verbatim*, *drawing a mind map containing main ideas*, and *summarizing main ideas in one's own words*.

We selected these strategies with four reasons in mind: First, we sought to design and implement an ecologically valid intervention. For this reason, we selected those strategies which the state curriculum recommends for regular fourth-grade German instruction in Bavaria (Bayerisches Staatsministerium für Unterricht und Kultus, 2000).

Second, students can effectively learn to use all three of these strategies during regular classroom instruction, and their use of these strategies can lead to improvements in finding main ideas and reading comprehension (for an overview, cf. National Institutes of Child Health and Human Development, 2000; Slavin, Lake, Chambers, Cheung, & Davis, 2009).

Third, as we designed our intervention for regular classrooms with children representing a wide spectrum of ability levels, we were interested in selecting strategies that are appropriate both for average readers (e.g., Bean & Steenwyk, 1981; Griffin, Malone, Kameenui, 1995) and for less advanced readers and students with learning disabilities (Kim, Vaughn, Wanzek, & Wei, 2004; Malone & Mastropieri, 1992).

Fourth, our choice of strategies also reflects findings indicating that teaching these strategies is particularly effective when they are taught in combination with one or more of the other steps covered in the seven-step cycle of self-regulated learning. Main-idea instruction is more effective when it is combined with self-monitoring than when it is presented by itself (Jitendra, Hoppes, & Xin, 2000; Malone & Mastropieri, 1992). Similarly, interventions combining instruction on finding main ideas in texts or on text comprehension strategies with goal setting is more effective than the same interventions without goal setting (Schunk & Rice, 1989; cf., however, Johnson, Graham, & Harris, 1997). Furthermore, evidence documents the superiority of teaching various metacognitive strategies in combination with text strategies in comparison to regular instruction or to teaching only text strategies (Mason, 2004; Souvignier & Mokhlesgerami, 2006). However, to our knowledge, there are no intervention studies in which students learn about a specific model of self-regulated learning and then—

<sup>1</sup> The subject is called *Heimat- und Sachunterricht* in Bavaria, Germany, and deals with basic aspects of everyday life, including topics from biology, geography, physics, health, and social sciences.



with substantive knowledge of the model—work systematically through the individual steps of the model.

While combining the presentation of a normative model with opportunities for practicing the application of the model's steps, our intervention design reflects these insights on effective self-regulated learning interventions. With these goals in mind, we designed a 7-week training program consisting of 2 informational weeks and 5 learning-cycle weeks. During the informational weeks, teachers introduce the seven-step cycle of self-regulated learning and the text reduction strategies mentioned previously. The knowledge presented during the informational weeks is then proceduralized in the five learning-cycle weeks. In other words, once students have understood the basic ideas behind the skills described in the seven-step cycle of self-regulated learning (during the two informational weeks), they then use the learning-cycle weeks to actually start developing these skills through practice with specific content (i.e., an expository text of the same length and difficulty level every day) and with the help of various learning materials. For example, participants receive learning journals (cf. Hübner, Nückles, & Renkl, 2010) with which they document their learning behavior, difficulties they encounter, and adjustments they make to their learning strategies (cf. the Method section). The learning journals and various other intervention materials help students to recognize the usefulness of the metacognitive and text strategies introduced in the intervention (Dignath & Büttner, 2008; Schunk & Rice, 1987). To facilitate this process, teachers give feedback and help the students to systematically draw connections between learning behavior and learning achievements (cf. the description in the Method section).

As we mentioned previously, meta-analyses indicate that teacher-led interventions are not as effective as researcher-lead interventions. However, in order to facilitate the transfer of the skills presented in the intervention to everyday practices throughout a child's school and homework activities, it is essential that classroom teachers lead the interventions. To increase the effectiveness of our teacher-led intervention, we placed an emphasis on the initial training of teachers prior to the administration of the program as well as on ongoing support during their implementation of the program. Before conducting the training program in their classrooms, teachers completed 2 full days of training. They also received extensive training materials and a teachers' manual designed to help them and their students avoid the sorts of barriers typically encountered in strategy instruction (cf. Kline, Deshler, & Schumaker, 1992). We also accompanied the administration of the entire 7-week training program (cf. Guskey, 1986).

In the present study, we examined whether the intervention as described leads to effects in students' self-reported preference for self-regulated learning, their ability to find main ideas in expository texts, and their overall reading comprehension. We compared three groups: students who receive regular instruction (REG), students who receive special instruction in text reduction strategies (TEXT), and students who receive instruction in text reduction strategies embedded in a training program focused on the seven-step cycle of self-regulated learning (SRL + TEXT). The comparison between the SRL + TEXT condition and the REG group shows the effect of the entire intervention approach compared with regular classroom instruction. This comparison is especially relevant from a practical perspective. The comparison between the SRL + TEXT condition and the TEXT condition shows the

additional benefit of teaching text reduction strategies within the context of a cycle of self-regulated learning. This comparison is especially relevant from a theoretical perspective.

In addition to a summative evaluation with pretest, posttest, and follow-up data collection, we also incorporated a process evaluation (cf. Stoeger & Ziegler, 2008a; Zimmerman, 2008). For the summative evaluation, all three groups of students filled out a learning preferences questionnaire and completed a standardized reading comprehension test at three points in time: before the intervention, directly after its conclusion, and then 11 weeks later. In our process evaluation, we observed whether the number of identified main ideas increased as students in the intervention groups worked on the daily expository texts.

In light of previous research, we expected an increase in the number of identified main ideas for both intervention groups over the course of the training program. We also expected that students in the combined intervention group (SRL + TEXT) would show a greater preference for self-regulated learning in comparison with the students in both other groups, both immediately after the intervention and in the follow-up test. Practicing metacognitive and text reduction strategies simultaneously appears to be more effective than only working on text strategies (cf. Dignath et al., 2008). We thus expected that the number of identified main ideas would increase more for the students in the combined intervention group (SRL + TEXT) over the course of the 7 weeks than it would in the group of students who only received the text strategy intervention (TEXT). As the focus of our intervention was mainly on basic text reduction strategies (cf. Cantrell, Almasi, Carter, Rintamaa, & Madden, 2010) and as standardized reading comprehension tests additionally measure other aspects of reading comprehension not explicitly covered in our intervention, we considered these tests to be transfer measures and expected to find weak to moderate effect sizes for our two intervention groups (cf. Souvignier & Mokhesgerami, 2006). We expected the best performance in the reading comprehension test for students in the combined intervention group, followed by students in the text-strategy-only intervention group, whom we expected to perform better than students in the regular instruction group.

## Method

### Participants and Design

Participants were 763 fourth-graders in 33 classrooms in urban, suburban, and rural areas in southern Germany. The students were on average 9.80 years old ( $SD = 0.43$ ); there was even gender distribution (48.89% girls). Among the participating students, 21.23% had a migration background (MB); that is, they themselves or at least one of their parents had not been born in Germany. The most common languages MB students learned as children were (in descending order) Russian, Turkish, Italian, Albanian, Serbian, and Bosnian. None of the students in our sample were rated by teachers as having difficulties understanding spoken German. Table 1 provides additional information about the MB students. As the linguistic backgrounds of the students varied but all were fluent German speakers, we had no a priori expectations about the effect of the students' migration status.

In our quasi-experimental study (Gliner, Morgan, & Leech, 2009), students in intact classrooms were recruited via the local



Table 1  
*Demographic Information by Treatment Condition*

Demographic information	Condition			
	SRL + TEXT ( <i>n</i> = 229)	TEXT ( <i>n</i> = 268)	REG ( <i>n</i> = 266)	Total ( <i>n</i> = 763)
Mean age in years ( <i>SD</i> )	9.89 (0.44)	9.80 (0.43)	9.74 (0.40)	9.80 (0.43)
Percentage of girls	48.03	50.75	47.74	48.89
Percentage of MB students (overall)	38.86	8.58	18.80	21.23
Percentage of MB students who				
Were not born in Germany	19.77	13.04	22.00	19.50
Use German as their primary language at home	47.20	78.26	66.00	57.41
Speak German at home at least sometimes	94.38	95.65	96.00	95.06

*Note.* SRL = self-regulated learning; TEXT = text reduction strategies; REG = regular classroom instruction; MB = migration background (the student and/or at least one of his or her parents were not born in Germany).

education authorities, who also gave us permission to conduct this study. The local education authorities offered all fourth-grade teachers in their district the opportunity to participate in an evaluation study of a classroom-based text-strategy training program as part of their professional development requirements. We semi-randomly assigned interested teachers, all of whom were certified elementary school teachers with at least 10 years of teaching experience to the three conditions (two intervention conditions and one regular instruction condition). When more than one teacher was participating at the same school, we assigned these teachers either to the same intervention condition or to the regular instruction condition, such that teachers were not aware that different versions of the program were being implemented. The teacher sample represented a total of 22 schools, with SRL + TEXT teachers distributed across eight of the schools and TEXT teachers distributed across nine of the schools. Four of the REG teachers taught in the same school as one of the SRL + TEXT teachers, and eight REG teachers taught in schools where there were no intervention-condition teachers. Teachers who were assigned to the regular instruction condition (under the pretense that we had a maximum number of participants and had raffled off the spots) were given the chance to receive the training materials after the evaluation study ended, and we promised them preferential admission to future workshops. At the end of the study, we debriefed all teachers about the study design and offered them feedback on the results of students in their own classrooms. Teachers' and students' participation in the evaluation study was voluntary, and both participants and their parents consented to participation. Teachers also informed the students' parents about the program.

We implemented a pretest (Time 1, or T1), posttest (T2), and follow-up test design (T3) with three conditions: In the final sample, nine<sup>2</sup> classrooms participated in the full training condition, practicing both self-regulated learning and text reduction strategies (SRL + TEXT). Twelve classrooms participated in the text-strategy-only condition (TEXT). The students in this condition received the same training as the full training group, but without the specific self-regulation components of the training. Students from 12 additional classrooms received regular instruction (REG). Table 1 shows our sample's demographic information by treatment condition. The evaluation of our study included two aspects: We conducted a summative evaluation for all three conditions at three measuring points with the help of standardized reading tests and questionnaires; we also carried out a process evaluation in the two

training conditions to evaluate the students' progress in finding main ideas in daily texts over the course of the training.

## Procedure

**Teacher workshops.** Before implementing one of the two versions of the training program (SRL + TEXT or TEXT, see later detailed description), each group of intervention-condition teachers attended a workshop designed to prepare them for administering their respective version of the training program in their classrooms. As teachers were to conduct evaluations in their classrooms themselves, they also learned how to administer the measurement instruments. Teachers in the regular instruction condition (REG) only learned how to administer the measurement instruments. The workshops were held by the first two authors of this report.

The 2-day workshop for the teachers in the SRL + TEXT condition covered theoretical information on text reduction strategies and self-regulated learning on the first day and the specific training program on the second day. Teachers received training materials for their students and discussed how they would administer them in their classrooms. They also received a teachers' manual documenting the concepts covered in the workshop and containing checklists of the materials to be covered on each day of the program (cf. Stoeger & Ziegler, 2008b).

As teachers in the TEXT condition did not learn about self-regulated learning, their workshop lasted only 1 day. These teachers received exactly the same instruction and material on text reduction strategies as teachers in the SRL + TEXT condition.

**Instruction in the three intervention conditions.** Instruction was delivered by fourth-grade classroom teachers during regular classroom hours. As the expository texts used in both intervention conditions dealt with topics from the natural sciences, the training was conducted mainly during reading instruction and instruction in basic science. The students in the regular instruction condition received a comparable amount of curriculum-based instruction in reading and basic science. As some classrooms with regular instruction were from the same schools as the training classrooms, we asked teachers in these classrooms not to employ any of the

<sup>2</sup> Originally, there were 12 participating classrooms in each of the three conditions. In the SRL + TEXT condition, three classroom teachers from one school decided on short notice and for reasons unrelated to the training program not to participate in the program.

material provided by us for the training classrooms during the study, but to teach their students as they normally would.

Teachers started administering their training program at dates scheduled shortly after the respective workshops. We provided all teachers of all three groups with contact information so that they could contact the first two authors of this report in the event that they were to have further questions regarding the implementation of the respective training program or the evaluation. Teachers in the intervention conditions could also contact their participating colleagues from the same group. Four (SRL + TEXT) or 3 weeks (TEXT) into the training program, we met with teachers in each of the intervention conditions in order to discuss practical issues of administering the program and to answer questions.

**Training program in the SRL + TEXT condition.** Classroom teachers in the SRL + TEXT condition implemented a 7-week program in which students practiced text reduction strategies as an integral part of self-regulated learning exercises (Stoeger & Ziegler, 2008b). The program included daily activities for regular classroom instruction and for homework. By completing the program, the students systematically practiced all phases of the cycle of self-regulated learning described in the Introduction section.

The training program consisted of 2 informational weeks and, thereafter, 5 learning-cycle weeks. During the informational weeks, students spent approximately 45–60 min of instruction time per day on the training program. During the learning-cycle weeks, the time spent on the training varied between approximately 40 min on Tuesdays, Wednesdays, and Thursdays and approximately 60 min on Mondays and Fridays.

During the first informational week, students learned why it is important to understand texts, what main ideas are, how they can identify them in expository texts, and how they can differentiate between main ideas and less important passages. Students received a one-page summary on how to identify main ideas; they were encouraged to refer to this summary throughout the program whenever they felt the need to do so. Teachers also presented and modeled three reduction strategies that are useful for identifying and displaying main ideas: (a) underlining and copying main ideas verbatim, (b) drawing a mind map containing main ideas, and (c) summarizing main ideas in one's own words. Students received a one-page summary on each strategy and were given the opportunity to practice each strategy on a short expository text (approximately 200–240 words).

During the second informational week, teachers introduced the self-regulated learning cycle by Ziegler and Stoeger (2005). For the students, the cycle was called the *learning circle* and was illustrated with cartoon-style pictures of Zumpel the Mouse who described all seven phases of the circle as a first-person narrator. Using this instructional material, students created their own learning circles and hung them up at home. Their hand-made learning-circle illustrations as well as the illustrations provided in the training program materials were meant to ensure that students would have frequent and easy access to visualizations of the learning circle and its individual phases while working through the training program. Teachers also used the second informational week to discuss the phases of self-regulated learning with their students; they used various examples drawn from everyday situations such as completing homework or practicing a certain sports skill. At the end of the second informational week, teachers provided their students with information on effective goal setting and

discussed common goal-setting mistakes with their students. As students should become aware of the relationship between using learning strategies and achieving goals and as this is a very demanding task for fourth graders, we asked students to set relatively simple quantitative outcome goals. Finally, teachers informed their students about the structure of the training program planned for the upcoming weeks.

During the following weeks, the learning-cycle weeks, the students repeatedly and consciously worked through all phases of the learning cycle. Every school day, students were to read an expository text about a topic from the natural sciences (e.g., fungi and mushrooms; rainbows; desert plants; blood) and then to identify the 10 main ideas. The texts were designed especially for use in the training program and to adhere to the following criteria: Each text was about 420 words long and contained 10 main ideas as well as several distractor sentences (see online supplemental material for a sample text). All texts were of a comparable difficulty level. The texts received a mean score of 69.16 ( $SD = 3.73$ ) on the German version of the Flesch readability index (Amstad, 1978), which corresponds to the difficulty level found in fifth-grade textbooks. These design criteria were set to ensure (a) that the texts would offer all students—including strong readers—the best possible chance of benefiting from applying, monitoring, and adjusting their strategy use and (b) that all students would be able to establish a clear connection between improved strategy use and better results. During the learning-cycle weeks, students kept a structured learning journal that accompanied them as they progressed through the learning cycle.

At the beginning of each learning-cycle week, students set a specific outcome goal for themselves that specified how many main ideas (10 being the maximum) per daily text they aspired to find. The students were encouraged to set goals for themselves that were challenging but achievable. They noted their goals in their learning journal, and they also wrote down what strategy they planned to use in order to achieve their goal. During learning-cycle Weeks 1–3, one of the three previously introduced text strategies for identifying and displaying main ideas was prescribed by the program per week: underlining and copying verbatim for the first learning-cycle week, mind mapping for the second, and summarizing for the third. This way, all students had the opportunity to practice each strategy systematically. In the remaining 2 learning-cycle weeks (learning-cycle Weeks 4 and 5), students chose strategies that they felt had been particularly helpful during the previous weeks and/or strategies from which they felt they could profit from continued practice of their effective implementation. During each of the 5 weeks, the students used their journals to keep track of how exactly they planned to use the strategy they were focusing on, how their monitoring worked, and what strategy adaptations they made.

In order to help students in the SLR + TEXT classrooms better understand the text strategies introduced during the first training week in the context of self-regulated learning, we incorporated an additional, story-based reading activity into learning-cycle Weeks 1, 2, and 3. We prepared four age-appropriate stories written in a more informal style in which the cartoon character Zumpel the Mouse served as a model of self-regulated learning use. In reading these “self-regulated learning stories,” the students accompany Zumpel as the mouse works on the aforementioned strategies: Zumpel self-assesses the learning process (Text 1) and then tries



out, monitors, and adjusts the underlining strategy (Text 2), the mind mapping strategy (Text 3), and the summarizing strategy (Text 4).

The students received one expository text per school day. They read the daily text silently and then had the opportunity to ask their peers and teacher about unfamiliar words. Then, before taking the text home and working further with it, they noted in their learning journal how many main ideas they thought they would find in that text (10 being the maximum number). At home, they used that week’s strategy to identify and display the main ideas in the text. Students spent between 20 and 30 min on this homework assignment. Right after having finished this part of their homework assignment, they evaluated how well their strategy worked on that day and wrote down in their learning journal how they wanted to improve their strategy use the next day. The next day, the homework assignment was discussed in class. Teachers based this discussion on the sample solutions they had received as part of the teachers’ manual. The students noted in their learning journal how many of the main ideas they actually found. In a teacher–class dialogue, the teacher addressed the connection between strategy use and outcome. Students were encouraged to use their experience with the text from the previous day to improve their strategy when working on the next text.

Each Friday, Thursday’s homework assignment was discussed first. Then, the students worked on a new text during classroom instruction. After discussing results and strategy use for this new text, the teacher initiated a discussion about learning behavior, strategy use, and results in the current week. We integrated appropriate prompts into the students’ learning journals to help facilitate this reflection process. The students thus also took time during classroom instruction on Fridays to summarize the current week in their journals. Based on this summary, teachers discussed the learning behavior with their students and how they could use their experience from this week to improve their learning behavior in the following week.

**Training program in the TEXT condition.** Teachers in the TEXT condition used the same materials and methods as teachers in the SRL + TEXT condition with one exception: They did not employ the materials on or make explicit references to self-regulated learning. As the TEXT-condition teachers did not introduce the concept of self-regulated learning to their students (Informational Week 2 in the SRL + TEXT condition), the duration of the TEXT-condition training program was reduced to 6 weeks. During an informational week, students in the TEXT condition learned—as did the students in the SRL + TEXT condition—why it is important to understand texts, what main ideas are, how they can identify them in expository texts, and how they can differentiate main ideas from less important passages. They also received the one-page summary on how to identify main ideas. Teachers in the TEXT condition also introduced and modeled the same three text reduction strategies used in the SRL + TEXT condition.

Then, during the subsequent five practice weeks, students applied the strategies to one expository text per school day by working to identify the 10 main ideas within each text. As in the SRL + TEXT condition, teachers discussed the correct solutions of this homework assignment with their students. However, students were not encouraged to use any self-regulated learning strategies. Students in the TEXT condition neither read self-

regulated learning stories nor kept learning journals. Table 2 shows the two intervention conditions in comparison.

**Instruction in the REG condition.** Students in the REG condition received regular classroom instruction in reading and basic science in accordance with the curriculum. The curriculum explicitly lists the use of text strategies such as underlining, making graphic representations, and summarizing as part of the reading instruction and summarizing basic scientific texts as part of the basic science instruction. Moreover, the legally binding state curriculum of the region where the study was conducted explicitly encourages teachers to emphasize self-regulated learning as the basis for lifelong learning and as a means of transferring more responsibility for the learning process onto the students. Within the confines of the curriculum, teachers in the regular instruction conditions could adjust their teaching to the needs of their students. Students spent between 20 and 30 min on their reading and basic science homework assignments each day.

Table 2  
*Two Intervention Conditions in Comparison*

SRL + TEXT	TEXT
Informational weeks	Informational weeks
Week 1	Week 1
Why understand texts	—
How to find main ideas	—
How to use text reduction strategies	—
Week 2	Week 2
Self-regulated learning	Why understand texts
—	How to find main ideas
—	How to use text reduction strategies
Learning-cycle weeks	Practice weeks
Daily tasks	Daily tasks
Reading	Reading
Read text	Read text
Use text reduction strategy	Use text reduction strategy
Find 10 main ideas	Find 10 main ideas
SRL	
Self-assessment	
Strategy monitoring	
Strategy adjustment	
Outcome evaluation	
Weekly tasks (Weeks 3–7)	
SRL	
Goal setting	
Strategic planning	
Outcome evaluation (reflection)	
Week 3	Week 3
Underlining	Underlining
SRL stories	—
Week 4	Week 4
Mind mapping	Mind mapping
SRL story	—
Week 5	Week 5
Summarizing	Summarizing
SRL story	—
Week 6	Week 6
Applying a text reduction strategy of choice	Applying a text reduction strategy of choice
Week 7	Week 7
Applying a text reduction strategy of choice	Applying a text reduction strategy of choice

Note. SRL = self-regulated learning; TEXT = text reduction strategies.

**Treatment fidelity.** All teachers indicated in their checklists that they used all training materials with the exception of one teacher who skipped one text due to a school activity day. The student training materials, which we collected from the students after the training programs were over, also suggest that the training programs were delivered as intended. Missing data in the student materials are in the range we expected for a practice period of 5 consecutive weeks. From this evidence and from more information collected in personal communication with teachers, we concluded that the interventions were implemented as intended.

**Program evaluation.** The testing sessions for the summative evaluation were scheduled during regular classroom hours in the week before the training started (T1), in the week after it concluded (T2), and another 11 weeks later (T3). The sessions were led by trained research assistants or by the specially trained classroom teachers.

At T1, students filled out the questionnaire on their preference for self-regulated learning during one 35-min testing session and completed the reading comprehension test and questions on demographic information in another testing session that lasted 25 min. At T2 and T3, the testing sessions lasted 35 min (questionnaire) and 75 min (reading comprehension test), respectively. To ensure comparable testing conditions, teachers and research assistants followed a detailed manual and read out instructions verbatim. The instrument for the process evaluation was integrated into the training material in both intervention conditions. An overview of our measurement schedule is provided as online supplemental material, Table S1.

## Measures

**Measures used in the summative evaluation.** We measured the preference for self-regulated learning at T1, T2, and T3 with the 28 items of the Fragebogen Selbstreguliertes Lernen–7, or FSL–7 [Questionnaire of Self-Regulated Learning–7] by Ziegler, Stoeger, and Grassinger (2010). The FSL–7 is based on Ziegler's and Stoeger's (2005) seven-step cyclical model of self-regulated learning. In the questionnaire, four school-relevant situations are described briefly: studying for school, preparing for the upcoming school year during the summer holidays, preparing for an in-class test, and catching up on school work after an illness. In each situation, the students are asked to indicate their preferred method of learning for each of the seven steps of self-regulated learning (self-assessment, goal setting, strategic planning, strategy implementation, strategy monitoring, strategy adjustment, and outcome evaluation) by choosing one of three alternatives: self-regulated, externally regulated, or impulsive learning. The following is a sample item (Situation 1, Step 2: Goal setting):

How do you study for school? (a) I set a fixed goal for myself describing what and how much I want to study [self-regulated learning]; b) My teacher or parents should tell me which goal I should set for myself [externally regulated learning]; c) When studying, I don't set a specific goal for myself. I can rely on my intuition [impulsive learning behavior].

In the present study, the research assistant or the classroom teacher read the four situations and the response alternatives out loud, ensuring that everyone, including weak readers, could complete the questionnaire both accurately and quickly.

In the present study, we restricted our interest to the preference for self-regulated learning. Therefore, we calculated an overall score by counting the frequency with which a student chose self-regulated learning and divided it by the number of items answered. For ease of understanding, the scores are reported as percentages. For example, a student who chose the self-regulated learning option in 13 out of the 28 items would be given a score of 46.43%. The internal consistency came to .85 at T1, .91 at T2, and .93 at T3.

At T1, we measured reading comprehension with the text comprehension section of the Ein Lesetest für Erst-bis Sechstklässler, or ELFE 1–6 [Reading Test for First to Sixth Graders], by Lenhard and Schneider (2006). In this section of the ELFE 1–6, students have 7 min to read 13 short texts (15–56 words) and answer a total of 20 multiple-choice questions. According to the authors of the test, students require different levels of reading skills to answer the questions. The skills are: finding information (five items), intersentential reading (eight items), and inferential reading (seven items). For the purpose of this study, we calculated the overall reading score (range: 0–20 points). Cronbach's alpha came to .82 in our sample.

We had originally planned to use the ELFE test at T2 and T3 as well. However, as we encountered unexpected ceiling effects at T1 (39.20% of all students had a score of at least 18 out of 20 points, and 11.33% of all students had a perfect score of 20 points), we decided to use a different, more difficult test at T2 and T3 that was designed to assess similar aspects of reading and text comprehension. We employed the text comprehension section of the Hamburger Lesetest für 3. und 4. Klassen, or HAMLET 3–4 [Hamburg Reading Comprehension Test for Grades 3 and 4] by Lehmann, Peek, and Poerschke (2006), using Version A at T2 and Version B at T3. Time constraints prevented us from employing both the ELFE and the HAMLET.

The text comprehension section of each HAMLET version comprises 10 texts: five expository texts, three so-called functional texts (e.g., recipes and timetables), and two narrative texts; the text length varies between 57 and 592 words. The test was administered in two parts: Students had 25 min to work on Texts 1–4, and after a 5-min break, they had another 40 min to work on Texts 5–10. Students were asked to answer four multiple-choice questions per text. According to the test's authors, students require different levels of reading skills to answer the questions. The skills are: simple finding of information (nine items in Version A, eight items in Version B), targeted finding of information (nine items in Version A, 10 items in Version B), combining/reconstructing (14 items in both Versions A and B), and connecting/infering (eight items in both Versions A and B). For the purpose of this study, we calculated the overall reading score (range: 0–40 points). Cronbach's alpha came to .90 at T2 and .92 at T3 in our sample.

**Measure used in the process evaluation.** Students in both the SRL + TEXT and the TEXT conditions were asked to identify the 10 main ideas in each of the 25 texts provided throughout the course of the training program (see previous section "Training program in the SRL + TEXT condition" for details). After the end of training, we collected all of the students' training materials. Trained research assistants checked the number of correctly identified main ideas in each text (range: 0–10 main ideas), using a list of the correct main ideas for each text as a reference. After completing this rating process, we returned the training materials to the stu-



dents. As a measure for the process evaluation, we used the weekly average of the number of correctly identified main ideas, resulting in five values per student.

Sample Drop-Out and Missing Data

In terms of the summative evaluation, 13 students (1.7% of the sample) missed the reading test at T1, 26 (3.4%) at T2, and 30 (3.9%) at T3. Of these students, one missed the reading test both at T1 and T3, and seven missed the reading test both at T2 and T3. Eleven students (1.4% of the sample) missed the questionnaire about the preference for self-regulated learning at T1, 25 (3.4%) at T2, and 36 (4.7%) at T3. Of these students, one missed the questionnaire at both T1 and T3, and five both at T2 and T3.

To handle missing data appropriately, we used state of the art methods (cf. Graham 2009; Schafer & Graham, 2002). As the program that we chose for the inferential analyses for the summative evaluation, HLM (Hierarchical Linear and Nonlinear Modeling software) Version 6.08 (Raudenbush, Bryk, Cheong, & Congdon, 2011), applies listwise deletion methods even if the full-information maximum-likelihood estimation (FIML) is chosen for regular two-level analyses, we used multiply imputed data sets for all inferential analyses with HLM (for all details regarding HLM analyses, see section “Overview of Statistical Procedures” in Results). A discussion of methods for multiple imputation of multilevel data is beyond the scope of this article but can be found, for example, in van Buuren (2011). We used the WinMICE software (Jacobusse, 2005) to generate five data sets under the hierarchical linear model. WinMICE makes use of a nested Gibbs sampler to estimate the parameters of the multilevel model for individual variables. We then analyzed the five sets simultaneously with HLM.

We received training materials from 476 of the 497 students in both training groups (221 SRL + TXT, 255 TXT) for use in the process evaluation; 233 students completed all texts, 61 students missed only one text, 157 students missed two to seven texts, and 25 students missed eight to 13 texts. Data of all students were included in further analyses. In terms of the different texts, there were between 2.9% and 22.1% missing data per text, with missing data below 10% for the first 18 texts and over 20% for only one of the texts in the final week of the training. To ensure consistency with our summative evaluation, we multiply imputed the missing data for the number of correctly identified main ideas with the

WinMICE software. We then analyzed the five imputed data sets simultaneously with HLM.

Results

Descriptives and Zero-Order Correlations

Table 3 shows descriptive statistics, proportions of between-classroom variance (the intraclass correlation, or ICC), and bivariate Pearson correlations for all measures used in the summative evaluation. The ICC indicates “the proportion of variance in the outcome that is between groups” (Raudenbush & Bryk, 2002, p. 36) rather than between individuals. Students chose self-regulated learning as their preferred approach to learning for slightly more than one third of all FSL-7 items. The rather large standard deviation indicates large differences between students. Right after the training, a small portion (7.57%) of the variance was located between classrooms, rising to a medium portion (16.23%) at follow-up. Students scored on average 15.57 points in the ELFE reading comprehension test, which is slightly higher than fourth graders in the norm sample (cf. Lenhard & Schneider, 2006). In the HAMLET reading comprehension tests, students also scored slightly better than students in the norm sample (cf. Lehmann et al., 2006). For both data-collection points, the proportion of variance between classrooms in the HAMLET was small (7.48% and 3.78%). The measures of self-regulated learning at different data-collection points were correlated, as were the measures of reading comprehension at different data-collection points; the preferences for self-regulated learning and reading comprehension were not correlated. Table 4 contains means and standard deviations for the dependent variables, listed separately for each condition and data-collection point. Both original values and z-transformed values are presented.

Table 5 contains descriptive statistics for the process evaluation. Both training groups started with slightly more than six correctly identified main ideas on average in the first week, with a slight advantage for the students in the TEXT condition. Over the course of the training program, students in the SRL + TEXT condition increased the number of correctly identified main ideas from week to week, suggesting a linear increase, whereas the number of correctly identified main ideas seems to remain rather constant in the TEXT condition.

Table 3  
Descriptive Statistics, Proportions Between Classroom Variance, and Bivariate Pearson Correlations

Variable	Scale	<i>M</i>	<i>SD</i>	ICC (%)	1	2	3	4	5
1. Preference for self-regulated learning (T1)	0–100	35.99	20.89	—	—				
2. Preference for self-regulated learning (T2)	0–100	38.17	25.66	7.57	.61**	—			
3. Preference for self-regulated learning (T3)	0–100	39.64	28.86	16.23	.49**	.68**	—		
4. Reading comprehension (T1, ELFE)	0–20	15.57	3.51	—	.06	.06	.05	—	
5. Reading comprehension (T2, HAMLET A)	0–40	28.36	6.03	7.48	–.03	–.00	.03	.58**	—
6. Reading comprehension (T3, HAMLET B)	0–40	29.54	5.68	3.78	–.03	.02	.04	.56**	.68**

Note. ICC = intraclass correlation; T1 = Time 1; ELFE = Ein Lesetest für Erst- bis Sechstklässler [Reading Test for First to Sixth Graders; Lenhard & Schneider, 2006]; HAMLET = Hamburger Lesetest für 3. und 4. Klassen [Hamburg Reading Comprehension Test for Grades 3 and 4; Lehmann, Peek, & Poerschke, 2006].

\*\* *p* < .01, two-tailed.

Table 4  
*Descriptive Statistics per Condition and Point of Measurement*

Condition	Time 1			Time 2			Time 3		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
Original values									
Preference for self-regulated learning									
SRL + TEXT	228	38.15	21.53	222	44.47	27.79	216	49.43	30.49
TEXT	261	36.64	21.67	257	37.70	24.48	253	38.71	27.98
REG	263	33.48	19.30	259	33.14	23.74	258	32.34	25.94
Reading comprehension									
SRL + TEXT	225	15.42	3.76	218	27.83	6.41	213	28.92	6.10
TEXT	262	15.38	3.46	262	28.59	5.93	261	29.75	5.89
REG	263	15.90	3.34	257	28.58	5.80	259	29.83	5.04
z-transformed values									
Preference for self-regulated learning									
SRL + TEXT	228	0.10	1.03	222	0.25	1.08	216	0.34	1.06
TEXT	261	0.03	1.04	257	-0.02	0.95	253	-0.03	1.97
REG	263	-0.12	0.92	259	-0.20	0.93	258	-0.25	0.90
Reading comprehension									
SRL + TEXT	225	-0.04	1.07	218	-0.09	1.06	213	-0.11	1.07
TEXT	262	-0.05	0.98	262	0.04	0.98	261	0.04	1.04
REG	263	0.09	0.95	258	0.04	0.96	259	0.05	0.89

Note. SRL = self-regulated learning; TEXT = text reduction strategies; REG = regular classroom instruction.

### Preliminary Analyses

We used chi-square tests (for percentage data) and univariate analyses of variance (to compare means) to examine whether the three groups were comparable with regard to their demographic composition and their pretest scores. The groups did not differ significantly in their gender distribution ( $p = .75$ ; SRL + TEXT 48.03%, TEXT 50.75%, REG 47.74% female) but differed significantly in the proportion of students with migration background (MB;  $p = .00$ ; SRL + TEXT 38.86%, TEXT 8.58%, REG 18.80%; using the Bonferroni correction, all three pairwise comparisons showed significant differences) and with regard to the students' mean age ( $p = .00$ ; SRL + TEXT: 9.89 years; TEXT, 9.80 years; REG: 9.74 years; the Bonferroni post hoc test showed that only SRL + TEXT differed significantly from REG, but TEXT did not differ from the other two conditions). The groups differed in terms of preference for self-regulated learning ( $p = .04$ ; again the only significant difference in the Bonferroni post hoc test was between SRL + TEXT and REG) but did not differ signifi-

cantly in terms of reading comprehension pretest scores ( $p = .18$ ; for means and standard deviations, cf. Table 4). To control for these individual variables, we included them as covariates in all inferential analyses.

### Overview of Statistical Procedures

As we recruited students in intact classrooms for this study, we used hierarchical linear models (Raudenbush & Bryk, 2002) to analyze our data. This method takes into account the fact that students within a classroom are more similar to each other than are randomly selected students and estimates standard errors associated with the regression coefficients in an appropriate way. We conducted all analyses with the software package HLM (Version, 6.08; Raudenbush et al., 2011), using the FIML algorithm for model estimations.

**Summative evaluation.** For the summative evaluation, we were interested in assessing the program's effects on the students' preference for self-regulated learning and on their reading com-

Table 5  
*Descriptive Statistics for Number of Correctly Identified Main Ideas per Condition and Week*

Condition	Number of correctly identified main ideas														
	Week 1			Week 2			Week 3			Week 4			Week 5		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
Original values															
SRL + TEXT	220	6.05	1.78	219	6.10	1.77	218	6.72	2.07	218	6.72	1.99	220	7.30	1.78
TEXT	255	6.24	1.54	253	5.58	1.63	255	6.57	1.58	251	6.19	1.71	244	6.44	1.72
z-transformed values															
SRL + TEXT	220	-0.06	1.07	219	0.16	1.03	218	0.04	1.13	218	0.15	1.07	220	0.25	0.99
TEXT	255	0.05	0.93	253	-0.14	0.95	255	-0.04	0.87	251	-0.13	0.92	244	-0.23	0.96

Note. SRL = self-regulated learning; TEXT = text reduction strategies.



prehension immediately after the training and 11 weeks later. Therefore, we specified four sets of models (two dependent variables  $\times$  two time points) in which we regressed the respective individual outcome variable on individual (Level 1) and classroom (Level 2) predictors. As all students from one classroom were in the same training condition, we specified the training condition on the classroom level.

We calculated four models for each set. We used the unconditional model (Model 0) to calculate the intraclass correlation (ICC) of the outcome variable. In Model 1, we included all Level-1 covariates, namely, gender, age, MB, pretest score in the preference for self-regulated learning, and pretest score in reading comprehension. In a first step, we allowed intercepts and slopes to vary between classrooms. In a second step, we fixed slopes if they did not vary significantly ( $p < .10$ ) between classrooms or if the reliability of the variable was less than .10 (cf. Cheung & Keeses, 1990). This model served as a reference model for the other two models, which include Level-2 variables. In Model 2, we calculated the effects of the two training conditions after controlling for individual variables. To this end, we included two dummy variables, SRL + TEXT and TEXT, to specify a classroom's adherence to a certain treatment condition, making the REG group the reference group. Again, we let slopes vary freely first and fixed them if they did not vary between classroom or if reliability was low. We calculated Model 3 to account for the fact that classrooms and training conditions differed substantially in their proportions of MB students. For this reason, we added the proportion of MB students as a Level-2 covariate. Model 3 shows the training effects for classrooms with an average proportion of MB students.

We report fixed effects based on model estimation with robust standard errors. As the software package HLM Version 6.08 does not provide standardized beta coefficients, we standardized all continuous variables (measures of reading comprehension and preference for self-regulated learning) before entering them into the models for easier interpretation of effects. The intercept of the regression equations can now be interpreted as the mean for a male student without MB who is of average age and with an average preference for self-regulated learning and an average reading score; and the slope coefficients show by how much the dependent variable changes in terms of proportions of a standard deviation if a predictor changes by one unit.

**Process evaluation.** The process evaluation was conducted to examine whether students in both training conditions identified an increasing number of main ideas across the course of the training. We assumed that students in both training conditions would become more proficient as the training progressed and that the increase in the number of correctly identified main ideas would be greater for the students in the SRL + TEXT condition. After inspecting the descriptive statistics for both groups, we used a linear growth model (cf. Raudenbush & Bryk, 2002) to predict the weekly average number of correctly identified main ideas with the five time points per student on Level 1, students on Level 2, and classrooms on Level 3. We used the original metric (number of correctly identified main ideas) for the outcome variable (as opposed to z-standardized values) to allow for the modeling of actual growth. All continuous covariates were z-standardized. We coded the time points from 0 (Week 1) to 4 (Week 5) so that a coefficient of 0 for the linear time parameter yields the initial status, that is, the average number of correctly identified main ideas during the

first of the five learning-cycle weeks. The weekly growth rate (slope) is indicated by the value for the linear time parameter. In a manner similar to our approach in the summative evaluation, we modeled student characteristics as covariates on the student level (here, Level 2) and the training condition as a dummy variable on the classroom level (here, Level 3).

Also in a manner similar to the procedure in the summative analysis, we calculated four models. From the unconditional model (Model 0), we took estimates for the variance components in both the intercept (initial status; number of correctly identified main ideas) and the slope (weekly increase in the number of correctly identified main ideas). In Model 1, we included all student-level covariates, namely, gender, age, MB, pretest score in the preference for self-regulated learning, and pretest score in reading comprehension, both for the intercept and the slope. We allowed all covariates to vary between classrooms in a first step, but fixed them using the same criteria as in the summative evaluation in a second step. We also used this procedure for the two remaining models. Model 1 served as the reference model for Models 2 and 3, in which we included classroom variables. In Model 2, we included training condition as predictor on the classroom level. As we compared only the two training conditions with each other in this analysis, one dummy variable (SRL + TEXT) was sufficient. Thus, the TEXT group became the reference group in this analysis. Finally, we also controlled for the proportion of MB students per classroom in Model 3.

## Summative Training Effects

We present the results of the hierarchical regression analyses in two sections. First, we describe the results regarding the students' preference for self-regulated learning, both right after the training and 11 weeks later. Second, we present the results regarding the students' reading comprehension, again for both the posttest and the follow-up test.

**Preference for self-regulated learning.** The results of the two-level analyses of the preference for self-regulated learning are presented in Table 6, with posttest results in the upper half and follow-up results in the lower half. Model 1 serves as a reference model and contains only individual input variables. As expected, the preference for self-regulated learning at T1 is a strong predictor for the preferences for self-regulated learning both at T2 and T3. In addition, there is a trend indicating that girls' preference for self-regulated learning generally increased more from T1 to T2 than that of boys. We did not, however, find the same effect at T3. At T3, the preference for self-regulated learning was instead slightly higher for older than for younger students. When we introduced classroom-level predictors in Models 2 and 3, the values of the individual predictor variables remained roughly the same. In Model 2, we found a small effect of the combined training condition (SRL + TEXT) on the students' preference for self-regulated learning at the posttest and a medium effect at follow-up. As expected, there were no significant training effects on the preference for self-regulated learning for the TEXT condition. Inclusion of the training conditions as predictors in the model explained almost 40% of the classroom-level variance in Model 1 for the posttest and almost 35% for the follow-up test. Controlling for the proportion of MB students per classroom in Model 3 enhanced the effects of the SRL + TEXT training condition both

Table 6  
Results of the Two-Level Analyses for the Preference for Self-Regulated Learning

Variable	Model 1			Model 2			Model 3		
	$\beta$	SE	$R^2$	$\beta$	SE	$R^2$	$\beta$	SE	$R^2$
SRL at posttest (Time 2)									
Intercept	−0.05	0.06		−0.16	0.08		−0.17	0.08	
Level 1									
Pretest SRL	0.60*	0.03		0.60*	0.03		0.60*	0.03	
Pretest reading	0.04	0.03		0.03	0.03		0.03	0.03	
Gender female	0.09	0.06		0.09†	0.06		0.09†	0.06	
Migration	0.00	0.09		−0.04	0.09		−0.02	0.09	
Age	0.03	0.03		0.02	0.03		0.02	0.03	
Level 2									
SRL + TEXT				0.33*	0.11		0.39*	0.13	
TEXT				0.06	0.10		0.03	0.10	
Migration (agg.)							−0.07	0.05	
$R^2$ Level 1			37.53%			37.52%			37.52%
$R^2$ Level 2			—			39.47%			45.18%
Deviance			1778.00			1769.09			1767.37
SRL at follow-up (Time 3)									
Intercept	−0.04	0.07		−0.25	0.16		−0.27	0.11	
Level 1									
Pretest SRL	0.46*	0.03		0.46*	0.03		0.46*	0.03	
Pretest reading	0.04	0.03		0.03	0.03		0.03	0.03	
Gender female	0.07	0.06		0.08	0.06		0.08	0.07	
Migration	−0.04	0.07		−0.08	0.08		−0.04	0.08	
Age	0.10*	0.03		0.09*	0.03		0.09*	0.03	
Level 2									
SRL + TEXT				0.53*	0.15		0.68*	0.15	
TEXT				0.18	0.15		0.11	0.14	
Migration (agg.)							−0.14*	0.07	
$R^2$ Level 1			25.32%			25.15%			25.17%
$R^2$ Level 2			—			34.37%			41.17%
Deviance			1871.00			1859.91			1856.32

Note.  $N = 763$  students from 33 classrooms. Values for Level-1 variables are set in italics if slopes varied freely between classrooms. SRL = self-regulated learning; TEXT = text reduction strategies; agg. = aggregated (the proportion of migration background students per classroom was aggregated from individual student data on migration background status). Variance-explained statistics were computed from the variance components with the following equations:  $R^2_{\text{Level 1}} = (\sigma^2[\text{unconditional model}] - \sigma^2[\text{fitted model}]) / \sigma^2(\text{unconditional model})$ .  $R^2_{\text{Level 2}} = (\tau_{00} [\text{Model 1}] - \tau_{00} [\text{Model with Level-2 variables}]) / \tau_{00} (\text{Model 1})$ .

†  $p < .10$ . \*  $p < .05$ .

at posttest and follow-up. The proportion of explained classroom-level variance rose to over 45% at the posttest and to over 41% at the follow-up test.

**Reading comprehension.** Table 7 shows the results of the two-level analyses for reading comprehension. In Model 1, the pretest reading comprehension scores strongly predict reading comprehension scores both at the posttest and at the follow-up. By contrast, MB students scored significantly and considerably lower on the reading comprehension test at the posttest and at the follow-up, even though the pretest scores were controlled. In addition, younger students scored slightly better both at the posttest and at the follow-up test. Finally, girls achieved slightly better reading scores than boys at the follow-up. The values of the individual predictor variables changed very little when we introduced classroom-level predictors in Models 2 and 3. Introducing the training conditions in Model 2 did not unveil any training effects on reading comprehension. Neither the SRL + TEXT condition nor the TEXT condition had a positive effect on students' reading comprehension, and that is true for both the posttest and the follow-up test. The introduction of the intervention vari-

ables explained only a very small portion of the Level-2 variance in Model 1 (3.64% for the posttest and 0.55% for the follow-up test). However, when we added the proportion of MB students as class-level predictor, the effectiveness of the combined training program emerged. In that case, students in the SRL + TEXT condition scored significantly higher for reading comprehension than students in the other two conditions at posttest. At follow-up, the effect remained visible as a trend ( $p = .06$ ). In Model 3, some of the Level-2 variance in Model 1 is explained (21.27% for the posttest, 15.82% for the follow-up).

### Process Training Effects

The variance decompositions into within- and between-school components in Model 0 showed significant variation among children within classrooms and significant variation between classrooms both for initial status and for the weekly growth rate. For initial status, 16.34% of the variance was between classrooms, and for the weekly growth rate, 29.81%. The fact that classrooms differed more over the course of time than in initial status is not



Table 7  
Results of the Two-Level Analyses for Reading Comprehension

Variable	Model 1			Model 2			Model 3		
	$\beta$	SE	$R^2$	$\beta$	SE	$R^2$	$\beta$	SE	$R^2$
Reading at posttest (Time 2)									
Intercept	0.04	0.06		0.02	0.08		-0.00	0.08	
Level 1									
Pretest SRL	-0.01	0.03		-0.02	0.03		-0.01	0.03	
Pretest reading	<i>0.56*</i>	<i>0.37</i>		<i>0.56*</i>	<i>0.04</i>		<i>0.56*</i>	<i>0.04</i>	
Gender female	0.10	0.06		0.10	0.06		0.10	0.06	
Migration	-0.41*	0.10		-0.42*	0.10		-0.38	0.10	
Age	-0.05 <sup>†</sup>	0.03		-0.06	0.03		-0.06	0.03	
Level 2									
SRL + TEXT				0.06	0.12		0.19*	0.09	
TEXT				0.03	0.09		-0.05	0.09	
Migration (agg.)							-0.15*	0.06	
$R^2$ Level 1			42.22%			42.21%			42.30%
$R^2$ Level 2			—			3.64%			21.27%
Deviance			1744.08			1743.70			1736.27
Reading at follow-up (Time 3)									
Intercept	0.00	0.06		0.02	0.06		-0.04	0.07	
Level 1									
Pretest SRL	-0.06*	0.03		-0.06*	0.03		-0.05*	0.03	
Pretest reading	<i>0.57*</i>	<i>0.04</i>		<i>0.57*</i>	<i>0.04</i>		<i>0.57*</i>	<i>0.04</i>	
Gender female	<i>0.14<sup>†</sup></i>	<i>0.08</i>		<i>0.15<sup>†</sup></i>	<i>0.08</i>		<i>0.15<sup>†</sup></i>	<i>0.08</i>	
Migration	-0.32*	0.09		-0.32*	0.10		-0.25*	0.11	
Age	-0.05 <sup>†</sup>	0.03		-0.05 <sup>†</sup>	0.10		-0.05 <sup>†</sup>	0.03	
Level 2									
SRL + TEXT				0.04	0.08		0.13 <sup>†</sup>	0.07	
TEXT				0.03	0.08		-0.04	0.10	
Migration (agg.)							-0.12*	0.04	
$R^2$ Level 1			40.37%			40.37%			40.51%
$R^2$ Level 2			—			0.55%			15.82%
Deviance			1781.08			1780.81			1773.97

Note.  $N = 763$  students from 33 classrooms. Values for Level-1 variables are set in italics if slopes varied freely between classrooms. SRL = self-regulated learning; TEXT = text reduction strategies; agg. = aggregated (the proportion of migration background students per classroom was aggregated from individual student data on migration background status). Variance-explained statistics were computed from the variance components with the following equations:  $R^2_{\text{Level 1}} = (\sigma^2[\text{unconditional model}] - \sigma^2[\text{fitted model}]) / \sigma^2(\text{unconditional model})$ ;  $R^2_{\text{Level 2}} = (\tau_{00} [\text{Model 1}] - \tau_{00} [\text{Model with Level-2 variables}]) / \tau_{00} (\text{Model 1})$ .

<sup>†</sup> $p < .10$ . \* $p < .05$ .

surprising: As the classrooms were assigned to different treatment conditions, different growth rates were to be expected.

The results of the growth model analysis estimating the increase of correctly identified main ideas in both training conditions are displayed in Table 8. In Model 1, only individual student characteristics were included. Reading pretest scores and gender positively predicted initial status, meaning that students with higher reading test scores as well as girls identified more main ideas correctly in the first week of the training. None of the individual covariates significantly predicted the linear trend in the course of the training, although there was a very small trend showing that the number of correctly identified main ideas increased less in the course of the training for older students. Introducing classroom-level variables into the model in Models 2 and 3 did not appreciably change the values of the individual predictors. Model 2 shows that the number of correctly identified main ideas in the first week was not predicted by treatment condition, indicating no significant differences between the SRL + TEXT and the TEXT group at the start of training. For the slope, we found a small effect for the SRL + TEXT condition: For students in this group, the

number of correctly identified main ideas increased by roughly one third ( $0.10 + 0.21 = 0.31$ ) of a main idea per week; in the TEXT condition, on the other hand, the number of correctly identified main ideas increased by only one tenth (0.10) of a main idea per week. The model estimated that by Week 5, students in the SRL + TEXT condition identified an average of 1.24 main ideas more than in Week 1, whereas students in the TEXT condition identified, on average, only 0.40 main ideas more than in their first week of training. These results remained stable when we controlled for the proportion of MB students per classroom in Model 3. The training condition remained the sole significant predictor of the growth rate in the course of the training and explained almost 50% of the between-classroom variance in the slope.

## Discussion

The current study was conducted with two main aims. From a theoretical perspective, the purpose was to assess the additional benefit of teaching text reduction strategies embedded in a training program focused on a normative model of self-regulated learning

Table 8  
Results of the Three-Level Growth Analysis for Correctly Identifying Main Ideas

Variable	Model 1			Model 2			Model 3		
	<i>b</i>	<i>SE</i>	<i>R</i> <sup>2</sup>	<i>b</i>	<i>SE</i>	<i>R</i> <sup>2</sup>	<i>b</i>	<i>SE</i>	<i>R</i> <sup>2</sup>
Intercept (initial status)	5.78	0.20		5.76	0.20		5.79	0.26	
Level 2 (student)									
Pretest SRL	0.08	0.07		0.08	0.07		0.08	0.07	
Pretest reading	0.41*	0.08		0.41*	0.08		0.41*	0.08	
Gender female	0.50*	0.15		0.50*	0.15		0.50*	0.15	
Migration	-0.22	0.19		-0.23	0.20		-0.24	0.21	
Age	0.06	0.05		0.06	0.05		0.06	0.05	
Level 3 (classroom)									
SRL + TEXT				0.03	0.33		-0.04	0.25	
Migration (agg.)							0.05	0.22	
Slope (growth rate)	0.18	0.04		0.10	0.04		0.09	0.05	
Pretest SRL	0.02	0.02		0.02	0.02				
Pretest reading	0.00	0.03		0.01	0.02				
Gender female	0.01	0.04		0.01	0.04		0.01	0.04	
Migration	0.03	0.04		0.00	0.04		0.00	0.04	
Age	-0.02 <sup>†</sup>	0.01		-0.03 <sup>†</sup>	0.01		-0.03 <sup>†</sup>	0.01	
Level 3 (classroom)									
SRL + TEXT				0.21*	0.06		0.22*	0.08	
Migration (agg.)							-0.01	0.05	
<i>R</i> <sup>2</sup> Level 2 intercept			19.60%			19.61%			19.61%
<i>R</i> <sup>2</sup> Level 2 slope			1.64%			1.24%			1.26%
<i>R</i> <sup>2</sup> Level 3 intercept			—			0.00%			0.00%
<i>R</i> <sup>2</sup> Level 3 slope			—			48.89%			49.55%
Deviance			8065.17			8050.44			8050.17

Note. *N* = 2,380 time points from 476 students in 21 classrooms. All Level-2 and Level-3 predictor variables were fixed. SRL = self-regulated learning; TEXT = text reduction strategies; agg. = aggregated (the proportion of migration background students per classroom was aggregated from individual student data on migration background status). Variance-explained statistics were computed from the variance components with the following equations:  $R^2_{\text{Level } 2} = (\sigma^2_{\text{unconditional model}} - \sigma^2_{\text{fitted model}}) / \sigma^2_{\text{unconditional model}}$ .  $R^2_{\text{Level } 3} = (\tau [\text{Model 1}] - \tau [\text{Model with Level-3 variables}]) / \tau (\text{Model 1})$ .

<sup>†</sup>  $p < .10$ . \*  $p < .05$ .

that students systematically study and proceduralize. To this end, we compared one group of students who learned text reduction strategies while also working on a self-regulated learning training routine (SRL + TEXT) with a second group who completed a training program focused exclusively on text reduction strategies (TEXT). From a more practical perspective, the purpose of this study was to examine the benefit of teacher-led text-reduction-strategy interventions (SRL + TEXT and TEXT) compared with regular classroom instruction (REG).

Our results generally confirm the effectiveness of the SRL + TEXT intervention and the advantage of this combined intervention over the pure text reduction strategy intervention (TEXT) and over regular classroom instruction (REG). In particular, the following findings apply to the three dependent variables we studied: First, as expected, both intervention groups showed linear increases in the number of main ideas identified in expository texts over the course of the respective intervention. We observed greater increases among the students in the combined intervention group (SRL + TEXT) than among those in the text-reduction-strategy-only group (TEXT). During the final week of the intervention, children in the combined intervention group identified almost one main idea more per expository text than the children in the text-strategy-only intervention group. This finding is consistent with the results of meta-analyses indicating that children in grade school, in particular, do best when they have the chance to work on a combination of cognitive and metacognitive strategies (Dignath et al., 2008). Other studies with a comparable evaluation design

(e.g., Stoeger & Ziegler, 2008a) have also revealed continuous improvements across an entire 5-week span of training. But unlike earlier studies, our current study documented training improvements that did not decline at the end of the intervention. This result suggests that the students' grasp on the text reduction strategies continuously improved and that they were also sufficiently motivated to continue using these strategies through to the very end of the training phase.

Second, these effects on finding main ideas in texts only transferred to higher scores in standardized reading comprehension tests in classrooms with no more than an average proportion of MB students. Further analyses comparing only students *without* a migration background revealed treatment effects in the standardized reading comprehension test in the combined intervention condition (SRL + TEXT). Non-MB students in the SRL + TEXT group demonstrated better reading comprehension at the posttest (Cohen's  $d = 0.25$ ) than the non-MB students in the group with regular classroom instruction. This advantage remained at the follow-up measurement, although it became less substantial (Cohen's  $d = 0.10$ ).

Students with MB in this intervention group performed nearly as well at mastering the skill of finding main ideas as those without MB, but the MB children were largely unsuccessful at applying this skill in the new context of reading comprehension in the standardized tests. One possible explanatory factor might be vocabulary deficiencies, especially concerning specific terminology. For the daily texts that students in both intervention groups worked on, teachers explained difficult words to the children before they



began working on the texts. This was not the case, however, for the standardized tests. It seems plausible that students with MB may have had special deficits in their language skills and breadth of vocabulary (Baumert & Schümer, 2001; Heinze, Herwartz-Emden, & Reiss, 2007) that might have influenced reading comprehension (cf. Ouellette, 2006).

Third, as expected, study participants who worked on text reduction strategies in the context of the seven-step cycle of self-regulated learning (SRL + TEXT) demonstrated an increased preference for self-regulated learning immediately after the training. The study participants in both of the comparison groups (TEXT and REG), on the other hand, showed no such changes. The effect we observed for the combined intervention group (SRL + TEXT) increased again from the posttest to the follow-up measurement 11 weeks after the training.

For the combined intervention condition (SRL + TEXT), the preference for self-regulated learning increased even more for students with MB than for those without. When we compared the MB children in this group to the MB children in the regular instruction group, we observed a preference-rating increase from the first to the second measurement of Cohen's  $d = 0.50$ ; for the children without MB, the effect only came to Cohen's  $d = 0.10$ . Both the MB and non-MB children showed even stronger preferences for self-regulated learning at our follow-up 11 weeks after the training. In comparison to MB children in the regular instruction group, MB children in the combined intervention group showed an increase in the strength of their preference for self-regulated learning from the first to the third measurement of Cohen's  $d = 0.64$ ; for non-MB children, the same comparison yielded a value of Cohen's  $d = 0.30$ .

## General Conclusion

Taken together, we come to the conclusion that these findings add to our understanding of how to increase older elementary students' preference for self-regulated learning and how to help them improve their ability at finding main ideas within an ecologically valid learning setting (De Corte, 2000). A comparison of the combined intervention group (SRL + TEXT) with the text-strategy-only intervention group (TEXT) and with the group receiving regular instruction (REG) shows that practicing text reduction skills within the framework of a normative model of self-regulated learning provides an additional benefit for elementary school children.

The positive development of students in the combined intervention is likely a result of the fact that the intervention adheres to four factors that researchers have identified as being beneficial (e.g., Dignath & Büttner, 2008; Ramdass & Zimmerman, 2011; Schunk & Rice, 1987; Weinstein, Husman, & Dierking, 2000): (a) We introduced the students in the combined intervention condition (SRL + TEXT) to a normative model of self-regulated learning, and they practiced each of the steps described in the model over the course of several weeks with concrete content and concrete strategies; (b) the intervention took place in more than one setting, namely, during regular classroom instruction in basic science and reading and during homework; (c) the intervention included various illustrations of the benefit of self-regulated learning; and (d) over the course of several weeks, students received systematic

feedback regarding their learning behavior and the relationship between this behavior and their achievements.

The effect sizes for finding main ideas through the combined intervention are comparable to—and the effect sizes for preference for self-regulated learning are somewhat greater than—those reported for earlier teacher-led interventions (Dignath & Büttner, 2008). The effect sizes for text comprehension are, however, somewhat smaller than in other previous studies (e.g., Paris, Cross, & Lipson, 1984). This difference may reflect the fact that we used standardized tests rather than tests designed specifically for our study. Researcher-designed tests tend to require less transfer of skills from one domain to another (e.g., Kim et al., 2004). Nevertheless, the obtained training effect sizes are lower than those reported for researcher-led training programs in small group settings (e.g., Dignath & Büttner, 2008).

## Limitations and Future Directions

Finally, we mention a number of limitations of our study. A first concern is about the assessment of self-regulated learning. With our assessment of the number of main ideas participants found over the course of the training and with the standardized reading tests, we established objective criteria for assessing achievement. Due to economic constraints, however, self-regulated learning was only assessed with a questionnaire. Thus, we did not measure students' actual behavior but rather their subjective assessments of their own behavior (cf. Artelt, 1999, 2000). In the case of the students in the combined intervention group in particular, this assessment approach can lead to distortions since these students may, by learning about self-regulated learning, be more prone to providing answers that they perceive as being socially desirable. For this reason, students' learning journals should be systematically evaluated in future research (cf. Schmitz, Klug, & Schmidt, 2011). Doing so should offer more insight into self-regulatory behavior during the training phase and provide some indication of the extent to which the self-assessments made in response to the relatively general questions in the questionnaire correspond with the journal entries (e.g., regarding self-monitoring and strategy adaptation) during the intervention. This brings us to a second specific recommendation for future work in this area: As learning-journal entries also reflect subjective assessments of one's own behavior, it would be helpful to also assess students' learning behavior using other approaches such as a microanalytic assessment method (Cleary, 2011), a think-aloud method (Greene, Robertson, & Croker Costa, 2011), or an in-depth case study method (Butler, 2011).

Another limitation is the possible occurrence of a Hawthorne effect, in that teachers changed their teaching behavior because they knew that their classrooms were being studied, not because of the specific intervention they received. However, the occurrence of a Hawthorne effect would have not been confined to the intervention groups because teachers in the REG condition also knew that their classrooms were being studied, such that they could have tried to improve their regular instruction in their own ways, thus making it harder to see intervention effects.

A final concern lies in the fact that we did not explicitly monitor the instruction that the participating teachers carried out during the intervention. Teachers did fill out daily checklists on the materials they used and on the aspects of the intervention which they dealt with, and the results suggest that the interventions fulfill criteria of



high treatment integrity. However, we do not have systematic data about how much time teachers spent working on each topic, which methods they preferred (e.g., group work or direct instruction), or how didactically effective their instructional approach was. We also did not collect data on the teachers' attitudes about self-regulated learning or about the intervention. In future research in this area, investigators may be able to incorporate the use of trained observers and/or video recording. In addition, asking teachers about their attitudes toward self-regulated learning, the intervention they are involved in, and its actual execution as well as testing their knowledge of self-regulated learning should provide important information about the conditions under which an intervention can be most effective.

In summary, the results of this study as well as those of other studies offer reason to be optimistic that self-regulated learning can be successfully introduced and practiced during classroom instruction and homework (cf. Ramdass & Zimmerman, 2011; Stoecker & Ziegler, 2011). The transfer of newly acquired self-regulated-learning knowledge and its proceduralization into skills is best facilitated by a combination of various intervention modules that employ various contents (e.g., mathematics, expository texts, vocabulary lists) and strategies (e.g., time management, text strategies, rehearsal strategies) within the framework of a normative model. In the future, researchers will need to (a) examine the efficacy of individual intervention modules, (b) better understand the conditions under which these modules are effective, and (c) look for evidence of both the advantages and the concrete effect of sequentially introducing and practicing the individual modules.

## References

- Alexander, P. A., Graham, S., & Harris, K. (1998). A perspective on strategy research: Progress and prospects. *Educational Psychology Review*, 10, 129–154. doi:10.1023/A:1022185502996
- Amstad, T. (1978). *Wie verständlich sind unsere Zeitungen?* [How readable are our newspapers?]. Doctoral thesis, Universität Zürich, Switzerland.
- Artelt, C. (1999). Lernstrategien und Lernerfolg. Eine handlungsnahe Studie [Learning strategies and learning success: An action-oriented study]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 31, 86–96. doi:10.1026//004637.31.2.86
- Artelt, C. (2000). *Strategisches Lernen* [Strategic learning]. Münster, Germany: Waxmann.
- Baumert, J., & Schümer, G. (2001). Familiäre Lebensverhältnisse, Bildungsbeteiligung, und Kompetenzerwerb [Family environment, education participation, and skill acquisition]. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, . . . M. Weiß (Eds.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 323–407). Opladen, Germany: Leske + Budrich.
- Bayerisches Staatsministerium für Unterricht und Kultus. (2000). *Lehrplan für die bayerische Grundschule* [Curriculum for Bavarian elementary school]. Munich, Germany: Maiß. Retrieved from Staatsinstitut für Schulqualität und Bildungsforschung (ISB) at [www.isb.bayern.de/schulartspezifisches/lehrplan/grundschule/](http://www.isb.bayern.de/schulartspezifisches/lehrplan/grundschule/)
- Bean, T., & Steenwyk, F. (1984). The effect of three forms of summarization instruction on sixth graders' summary writing and comprehension. *Journal of Literacy Research*, 16, 297–306. doi:10.1080/10862968409547523
- Butler, D. L. (2011). Investigating self-regulated learning using in-depth case studies. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 346–360). New York, NY: Routledge.
- Cantrell, S. C., Almasi, J. F., Carter, J. C., Rintamaa, M., & Madden, A. (2010). The impact of a strategy-based intervention on the comprehension and strategy use of struggling adolescent readers. *Journal of Educational Psychology*, 102, 257–280. doi:10.1037/a0018212
- Cheung, K., & Keesee, J. (1990). Hierarchical linear modelling. *International Journal of Educational Research*, 14, 289–297. doi:10.1016/0883-0355(90)90039-B
- Cleary, T. J. (2011). Professional development needs and practices among educators and school psychologists [Special issue]. *New Directions for Teaching and Learning*, 2011(126), 77–87. doi:10.1002/tl.446
- Council of the European Union. (2002, July 9). Resolution of 27 June, 2002, on lifelong learning. *Official Journal of the European Communities*, pp. C163/1–C163/3.
- De Corte, E. (2000). Marrying theory building and the improvement of school practice: A permanent challenge for instructional psychology. *Learning and Instruction*, 10, 249–266. doi:10.1016/S0959-4752(99)00029-8
- Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students: A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning*, 3, 231–264. doi:10.1007/s11409-008-9029-x
- Dignath, C., Büttner, G., & Langfeldt, H.-P. (2008). How can primary school students learn self-regulated learning strategies most effectively?: A meta-analysis on self-regulation training programs. *Educational Research Review*, 3, 101–129. doi:10.1016/j.edurev.2008.02.003
- Gliner, J. A., Morgan, G. A., & Leech, N. L. (2009). *Research methods in applied settings: An integrated approach to design and analysis* (2nd ed.). New York, NY: Routledge.
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549–576. doi:10.1146/annurev.psych.58.110405.085530
- Greene, J. A., Robertson, J., & Croker Costa, L. J. (2011). Assessing self-regulated learning using think aloud methods. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 313–328). New York, NY: Routledge.
- Griffin, C. C., Malone, L. D., & Kame'enui, E. J. (1995). Effects of graphic organizer instruction on fifth-grade students. *Journal of Educational Research*, 89, 98–107. doi:10.1080/00220671.1995.9941200
- Guskey, T. R. (1986). Staff development and the process of teacher change. *Educational Researcher*, 15, 5–12. doi:10.3102/0013189X015005005
- Hattie, J. A., Biggs, J., & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. *Review of Educational Research*, 66, 99–136. doi:10.3102/00346543066002099
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77, 81–112. doi:10.3102/003465430298487
- Heinze, A., Herwartz-Emden, L., & Reiss, K. (2007). Mathematikkenntnisse und sprachliche Kompetenz bei Kindern mit Migrationshintergrund und zu Beginn der Grundschulzeit [The mathematics knowledge and language competency of children with a migration background at the beginning of primary school]. *Zeitschrift für Pädagogik*, 53, 562–581.
- Hübner, S., Nückles, M., & Renkl, A. (2010). Writing learning journals: Instructional support to overcome learning-strategy deficits. *Learning and Instruction*, 20, 18–29. doi:10.1016/j.learninstruc.2008.12.001
- Jacobus, G. (2005). *WinMICE user's manual*. Retrieved from [www.multiple-imputation.com](http://www.multiple-imputation.com)
- Jitendra, A. K., Hoppes, M. K., & Xin, Y. P. (2000). Enhancing main idea comprehension for students with learning problems: The role of a summarization strategy and self-monitoring instruction. *Journal of Special Education*, 34, 127–139. doi:10.1177/002246690003400302
- Johnson, L., Graham, S., & Harris, K. R. (1997). The effects of goal setting and self-instruction on learning a reading comprehension strategy: A study of students with learning disabilities. *Journal of Learning Disabilities*, 30, 80–91. doi:10.1177/002221949703000107
- Kim, A.-H., Vaughn, S., Wanzek, J., & Wei, S. (2004). Graphic organizers and their effects on the reading comprehension of students with LD: A



- synthesis of research. *Journal of Learning Disabilities*, 37, 105–118. doi:10.1177/00222194040370020201
- Kline, F. M., Deshler, D. D., & Schumaker, J. B. (1992). Implementing learning strategy instruction in class settings: A research perspective. In M. Pressley, K. R. Harris, & J. T. Guthrie (Eds.), *Promoting academic competence and literacy in school* (pp. 361–406). San Diego, CA: Academic Press.
- Lehmann, R. H., Peek, R., & Poerschke, J. (2006). *HAMLET 3–4: Hamburger Lesetest für 3. und 4. Klassen* [Hamburg Reading Comprehension Test for Grades 3 and 4]. Göttingen, Germany: Hogrefe.
- Lenhard, W., & Schneider, W. (2006). *ELFE 1–6: Ein Leseverständnistest für Erst- bis Sechstklässler* [ELFE 1–6: A reading comprehension test for students in Grades 1 through 6]. Göttingen, Germany: Hogrefe.
- Malone, L. D., & Mastropieri, M. A. (1992). Reading comprehension instruction: Summarization and self-monitoring training for students with learning disabilities. *Exceptional Children*, 58, 270–279.
- Mason, L. H. (2004). Explicit self-regulated strategy development versus reciprocal questioning: Effects on expository reading comprehension among struggling readers. *Journal of Educational Psychology*, 96, 283–296. doi:10.1037/0022-0663.96.2.283
- National Institute of Child Health and Human Development. (2000). *National Reading Panel: Teaching children to read. An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups* (NIH Publication No. 00–4754). Washington, DC: Government Printing Office.
- Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology*, 98, 554–566. doi:10.1037/0022-0663.98.3.554
- Paris, S., Cross, D. R., & Lipson, M. Y. (1984). Informed strategies for learning: A program to improve children's reading awareness and comprehension. *Journal of Educational Psychology*, 76, 1239–1252. doi:10.1037/0022-0663.76.6.1239
- Pressley, M., Graham, S., & Harris, K. (2006). The state of educational intervention research as viewed through the lens of literacy intervention. *British Journal of Educational Psychology*, 76, 1–19. doi:10.1348/000709905X66035
- Ramdass, D., & Zimmerman, B. (2011). Developing self-regulation skills: The important role of homework. *Journal of Advanced Academics*, 22, 194–218. doi:10.1177/1932202X1102200202
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. (Vol. 1 in *Advanced Quantitative Techniques in the Social Sciences Series*, 2nd ed.). Thousand Oaks, CA: Sage.
- Raudenbush, S. W., Bryk, A. S., Cheong, Y., & Congdon, R. (2011). *HLM Version 6.08: Hierarchical linear and nonlinear modeling* [Software]. Chicago, IL: Scientific Software International.
- Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanism of a neglected phenomenon. *Educational Psychologist*, 24, 113–142. doi:10.1207/s15326985ep2402\_1
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177. doi:10.1037/1082-989X.7.2.147
- Schmitz, B., Klug, J., & Schmidt, M. (2011). Assessing self-regulated learning using diary measures with university students. In B. Zimmerman & D. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 251–266). New York, NY: Routledge.
- Schunk, D. H., & Rice, J. M. (1987). Enhancing comprehension skill and self-efficacy with strategy value information. *Journal of Reading Behavior*, 19, 285–302. doi:10.1080/10862968709547605
- Schunk, D. H., & Rice, J. M. (1989). Learning goals and children's reading comprehension. *Journal of Literacy Research*, 21, 279–293. doi:10.1080/10862968909547677
- Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research*, 79, 1391–1466. doi:10.3102/0034654309341374
- Souvignier, E., & Mokhesgerami, J. (2006). Using self-regulation as a framework for implementing strategy-instruction to foster reading comprehension. *Learning and Instruction*, 16, 57–71. doi:10.1016/j.learninstruc.2005.12.006
- Stahl, N. A., Simpson, M. L., & Hayes, C. G. (1992). Ten recommendations from research for teaching high-risk college students. *Journal of Developmental Education*, 16, 2–10.
- Stoeger, H., & Ziegler, A. (2008a). Evaluation of a classroom based training to improve self-regulated learning in time management tasks during homework activities with fourth graders. *Metacognition and Learning*, 3, 207–230. doi:10.1007/s11409-008-9027-z
- Stoeger, H., & Ziegler, A. (2008b). *Trainingshandbuch selbstreguliertes Lernen II: Grundlegende Textverständnisstrategien für Schüler der 4. bis 8. Jahrgangsstufe* [Accompanying manual for a training of self-regulated learning II: Basic text strategies for fourth- to eighth-grade students]. Lengerich, Germany: Pabst.
- Stoeger, H., & Ziegler, A. (2011). Self-regulatory training through elementary-school students' homework completion. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 87–101). London, England: Routledge.
- van Buuren, S. (2011). Multiple imputation of multilevel data. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 173–196). New York, NY: Routledge.
- Weinstein, C. E., Husman, J., & Dierking, D. R. (2000). Self-regulation interventions with a focus on learning strategies. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulated learning* (pp. 727–747). San Diego, CA: Academic Press.
- Weinstein, C. E., & Mayer, R. E. (1986). The teaching of learning strategies. In M. Wittrock (Ed.), *Handbook of research on teaching* (pp. 315–327). New York, NY: Macmillan.
- Ziegler, A., & Stoeger, H. (2005). *Trainingshandbuch selbstreguliertes Lernen I: Lernökologische Strategien für Schüler der 4. Jahrgangsstufe Grundschule zur Verbesserung mathematischer Kompetenzen* [Accompanying manual for a training of self-regulated learning I: Resource strategies for fourth-grade elementary school students to improve math skills]. Lengerich, Germany: Pabst.
- Ziegler, A., Stoeger, H., & Grassinger, R. (2010). Diagnostik selbstregulierten Lernens mit dem FSL–7 [Assessing self-regulated learning with the FSL–7]. *Journal für Begabtenförderung*, 10, 24–33.
- Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, 81, 329–339. doi:10.1037/0022-0663.81.3.329
- Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, 25, 3–17. doi:10.1207/s15326985ep2501\_2
- Zimmerman, B. J. (2000). Attaining self-regulation. A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). San Diego, CA: Academic Press.
- Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, 45, 166–183. doi:10.3102/0002831207312909
- Zimmerman, B. J., Bonner, S., & Kovach, R. (1996). *Developing self-regulated learners: Beyond achievement to self-efficacy*. Washington, DC: American Psychological Association. doi:10.1037/10213-000

Received November 26, 2012

Revision received November 25, 2013

Accepted December 5, 2013 ■

# Can Babies Learn to Read? A Randomized Trial of Baby Media

Susan B. Neuman  
New York University

Tanya Kaefer  
Lakehead University

Ashley Pinkham  
University of Michigan

Gabrielle Strouse  
University of Toronto

Targeted to children as young as 3 months old, there is a growing number of baby media products that claim to teach babies to read. This randomized controlled trial was designed to examine this claim by investigating the effects of a best-selling baby media product on reading development. One hundred and seventeen infants, ages 9 to 18 months, were randomly assigned to treatment and control groups. Children in the treatment condition received the baby media product, which included DVDs, word and picture flashcards, and word books to be used daily over a 7-month period; children in the control condition, business as usual. Examining a 4-phase developmental model of reading, we examined both precursor skills (such as letter name, letter sound knowledge, print awareness, and decoding) and conventional reading (vocabulary and comprehension) using a series of eye-tracking tasks and standardized measures. Results indicated that babies did not learn to read using baby media, despite some parents displaying great confidence in the program's effectiveness.

**Keywords:** early literacy, reading, baby media, eye tracking, vocabulary

There has been an explosion of new media targeted to babies in recent years (Rideout & Hammel, 2006). Ignited by the 1995 release of *Brainy Baby*, followed by a deluge of other videos and DVDs such as *Baby Einstein*, educational media specially marketed to families of infants and toddlers has become big business, with the Baby Einstein brand alone selling over \$200 million worth of products (Robb, Richert, & Wartella, 2009). A substantial proportion of these new media claim to promote infants' cognitive development and vocabulary and feature testimonials and advertisements about how young children may benefit from these commercial products. Garrison and Christakis (2005), for example, reported that of the top 100 best-selling baby DVDs on Amazon.com, 76 claim to produce specific developmental benefits. According to recent reports, these claims appear to be reaching an increasingly receptive audience: The average 6-month-old is said to own at least four DVDs (Barr, Danziger, Hilliard, Andolina, &

Ruskis, 2009), with this figure almost doubling by the time infants are 18 months old (Linebarger & Vaala, 2010).

A small proportion of the makers of these baby media, however, go beyond promoting developmental benefits to argue for their product's ability to teach babies to read. Claiming that "all babies are Einsteins," the manufacturers of Intellectual Baby, for example, make the case that babies can learn to read beginning at age 3 months (Intellectual Baby, 2009). Similarly, recognizing that babies are "linguistic geniuses," the makers of Brill Baby ("kids are brilliant"; BrillKids, 2011) promote using their Little Readers and Little Musicians series starting at infancy. Other products, like *Your Baby Can Read* (www.ybcr.com; Titzer, 2010), claim that toddlers will be able to read *Charlotte's Web* and *Harry Potter* if regularly exposed to their program as early as age 3 months. "Teaching your baby to read is easy," one product claims (Doman & Doman, 2010). "It depends on the brain's ability to integrate its visual, auditory, linguistic and conceptual centers. And that speed depends a great deal on the myelination of the neuron's axons . . . the more myelin sheathes the axon, the faster the neuron can conduct its charge. In short, the earlier we can teach babies to read, the better" (Intellectual Baby, 2009).

Although skeptics abound (American Academy of Pediatrics, 1999), a small number of empirical studies have begun to examine this assumption, targeting their focus specifically on word learning. Vandewater (2011), for example, in a randomized experiment, reported that infants exposed to Baby Wordsworth for 1 month showed greater gains in receptive vocabulary at a 3-month follow-up compared with controls, although their assessment of gains relied on parent report rather than direct assessments with children. In contrast, Robb et al. (2009) found no greater gains in receptive and expressive vocabulary for 12- and 15-month-olds who viewed the same product for 6 weeks versus a nonviewing

---

This article was published Online First February 24, 2014.

Susan B. Neuman, Teaching and Learning Department, Steinhardt School of Culture, Education, and Human Development, New York University; Tanya Kaefer, Department of Education, Lakehead University; Ashley Pinkham, School of Education, University of Michigan; Gabrielle Strouse, Department of Applied Psychology and Human Development, University of Toronto.

This research was conducted at the University of Michigan. We gratefully acknowledge the Ready to Learn research team and the children and teachers who participated in the study.

Correspondence concerning this article should be addressed to Susan B. Neuman, Teaching and Learning Department, Steinhardt School of Culture, Education, and Human Development, New York University, 239 Greene Street, New York, NY 10003. E-mail: sbneuman@nyu.edu



control group. Similarly, DeLoache and her research team (DeLoache et al., 2010) reported no difference in vocabulary gains for 12- to 18-month-old children who viewed the same product several times a week over a month's time versus a control group.

Nevertheless, there are limitations to these studies that might be contested by program developers. For one, many of these products recommend a longer duration of treatment than multiple exposures over 1 month's time. Your Baby Can Read, for example, suggests that parents are likely to experience the benefits for their child only after 3–7 months of daily use (Titzer, 2010). Second, many top-selling products include more than DVDs; for example, Intellectual Baby includes flashcards and books that accompany the program's DVD. And third, to date, studies have only measured gains in receptive and expressive vocabulary development. *Reading proficiency*, as we later define, involves more than oral word learning.

Therefore, one could make a case that claims about babies' abilities to learn to read have not been subject to rigorous empirical testing. However, some might argue that the question is moot, given the reported difficulty that infants and toddlers have learning from screen media (Anderson & Pempek, 2005). For example, one reason young children may struggle to learn from video presentations is that they do not display dual representation, or the understanding that pictures and videos depict not only objects in and of themselves but also represent similar objects in their world (DeLoache, 2004). Development of dual representation helps support children's generalization and transfer of content from screen presentations to other situations. Linebarger and Vaala (2010) argued that learning and extending new vocabulary, in particular, may be especially dependent on dual representation, as this skill requires infants to simultaneously represent an object on screen both as its own entity and as a symbol for similar objects in other environments.

A substantial amount of research suggests that infants and toddlers may not learn information as readily from screen media as from a live situation, a phenomenon known as the *video deficit* (Anderson & Pempek, 2005). The deficit has been found using a variety of language outcomes, including phonetic learning (Kuhl, Tsao, & Liu, 2003), connecting actions with a novel word (Roseberry, Hirsh-Pasek, Parish-Morris, & Golinkoff, 2009), and identifying object labels (Krcmar, Grela, & Lin, 2007; Troseth, Strouse, Verdine, & Saylor, 2013). Nevertheless, despite young children's inefficient learning from video, studies that have compared learning from video instruction with no instruction have suggested that infants and toddlers are capable of learning information from video (e.g., Barr & Hayne, 1999; Strouse & Troseth, 2008), especially with supportive scaffolds such as co-viewing and repeated viewings (Barr, Muentener, Garcia, Fujimoto, & Chávez, 2007; Strouse & Troseth, 2013). In addition, Rice (1983) identified a number of supportive features of videos themselves including the use of predictable program formats, recasts, simple sentences, and slow rates of speech, some of which are prevalent in baby media products.

Despite varied research findings, many parents retain the belief that their children benefit from watching television. For example, a recent survey indicated that 40% of mothers of young children believe that their infants and toddlers are learning from screen time (Rideout, 2007). Therefore, it is conceivable that certain precursors of reading (e.g., phonological awareness) are developing but may not be apparent due to limitations in the methods traditionally used

to assess early learning. Consequently, this article was designed to take marketers at their word and to measure the effects of baby media using a more comprehensive model of reading. Conducting a year-long randomized controlled trial in which we compared families who used a baby media product with relatively high fidelity with control families, we examined the effects of baby media on babies' reading development.

## What Is Reading?

Critical to a study of reading development is the very definition of "what is reading?" We use the widely accepted definition of reading reported in government consensus documents as well as in a convergence of studies (National Early Literacy Panel, 2008; National Reading Panel, 2000). Reading is a complex cognitive process of decoding symbols for the intention of deriving meaning from print. Although it has its detractors, the "simple view" most succinctly characterizes the process (Gough & Tunmer, 1986): Reading with understanding is the product of decoding and comprehension, or the simple formula,  $R = D \times C$ . In actuality, however, this definition is hardly simple because it suggests a multiplicative effect: Decoding by itself is not reading; similarly, comprehension of words without the ability to unlock words into their constituent parts is not reading. Both must work in concert for individuals to be able to read with meaning.

This definition contrasts with those who might argue that identifying words in context such as "McDonald's" or "Stop" in stop signs (otherwise known as environmental print) are indicators of real reading (Goodman, 1984). However, two independent studies—Masonheimer, Drum, and Ehri (1984) and Stahl and Murray (1993)—have shown that while young children as early as age 2½ years could readily and accurately identify many logos, they could not read the embedded words when they were removed from the logos. Although these behaviors, sometimes described as *pseudo-reading* or *pretend reading* (Teale & Sulzby, 1989), may highlight children's interest or awareness of symbol systems in their environment, they do not constitute reading or the ability to read words accurately in isolation or in text with meaning.

Nevertheless, the development of reading ability is not an all-or-nothing phenomenon. Rather, researchers agree that there are developmental phases that emerge or change based on internal causes, such as developing cognitive or linguistic capabilities, and on external environmental conditions, such as the scaffolding kinds of adult activities that support its development. In fact, there is substantial agreement among theories of reading development (Chall, 1983; Ehri, 1979; Mason, 1980) that children move from contextual dependency to early decoding, where they become "glued to print" by processing letters and sounds in an effortful manner, ultimately to fluent and conventional reading. Such developmental theories can provide researchers with a basis for assessing development and for predicting what can be expected in the developmental path toward reading.

Consequently, our analysis of whether babies can learn to read was based on the premise that there are developmental skills that act as precursors to reading proficiency. Using this logic, we would presume that even if babies could not read with understanding as a result of the baby media program (i.e., as indicated by program developers' claims) they might at least exhibit some of their earlier skills that could accelerate later reading development.

Consistent with this thesis, therefore, in this article, we examine reading developmentally from the emerging to the consolidation phases of reading.

## Method

### Participants

Participants were 117 infants, ages 10–18 months ( $M = 14.25$  months), and their parents from a small Midwestern city and the surrounding county. These families were recruited through a variety of sources, including flyers distributed in the community (e.g., churches, day care centers), displays at community events (e.g., public libraries, farmers' markets), social networking (e.g., Facebook), and word of mouth. We used the following inclusion criteria for participation in the study: (a) Babies were born full-term (i.e.,  $> 37$  weeks gestation), (b) heard English as their primary language (i.e.,  $< 10\%$  exposure to languages other than English), and (c) did not have a history of vision, hearing, or cognitive disabilities. Together, our sample included 61 male and 56 female infants, representing 88% White, 3% African American, and 9% bi/multiracial groups.

Infants were randomly assigned to one of two conditions (treatment or control) within counterbalanced gender and age (three clusters: 10.0–12.9; 13.0–15.9; and 16.0–18.9 months) brackets. The distribution of infants in each condition did not vary across demographic characteristics (all  $ps > .05$ ). See Table 1 for demographic information.

The majority of infants were from middle-class, highly educated parents. Half of the families earned an annual income larger than \$76,000 and 16% earned less than \$30,000. Seventy-eight percent of mothers and 75% of fathers had at least a bachelor's degree. Ninety-two percent of parents were married (single: 6.6%; domestic partners: 1.1%).

Almost half of the infants were first born and attended some form of day care at the start of the study: 13% in their own home (e.g., looked after by a grandparent), and 34% went to a day care center). The quality of infants' home environments was relatively high: 11% of families got the maximum score (i.e., 45 points) on the Infant/Toddler Home Observation for Measurement of the Environment Inventory (Caldwell & Bradley, 2003). Another 83% of the families scored close to the maximum score (i.e., 39–44 points). Bayley scores in cognition, language, and social-emotional development were in the average range (Bayley, 2006).

### Intervention Materials

The intervention included a best-selling baby media product, *Your Baby Can Read*, sold at popular chain stores (e.g., Walmart, Target) as well as through online vendors. Marketed to children ages 3 months and older, it purports to teach babies how to read with fluency and comprehension within 3–6 months of regular use. The intervention is composed of five volumes or units; each volume includes a specific number of words to be learned, ranging from 20 to 27 words. In Volume 1, for example, babies are expected to learn 22 written words, ranging from two to eight

Table 1  
*Demographic Characteristics of Treatment and Control Children (N = 117)*

Characteristic	Treatment ( $n = 61$ )	Control ( $n = 56$ )	$p^a$
Age (in no. of months)			
Initial visit	14.28 (2.60)	14.21 (2.70)	.887
Final visit	22.04 (2.81)	21.46 (2.91)	.294
Gender (%)			.236
Male	57.4	46.4	
Female	42.6	53.6	
Ethnicity (%)			.194
White	83.6	92.9	
African-American	4.9	—	
Bi-/multiracial	11.5	7.1	
Language exposure (%)			.676
English only	81.7	78.6	
Multiple	18.3	21.4	
Siblings (%)			.849
Only child	47.5	43.6	
1	39.0	40.0	
2+	13.6	16.3	
Attends child care	58.3	44.6	.140
Parental education ( <i>Mdn</i> )	Bachelor's degree	Bachelor's degree	.283
Household income ( <i>Mdn</i> )	\$76,000–\$100,000	\$51,000–\$75,000	.825
Infant/Toddler HOME score	41.69 (2.74)	41.69 (3.29)	.215
Bayley Scales percentile			
Cognitive	56.36 (23.19)	51.28 (24.43)	.251
Language	42.07 (25.70)	40.38 (23.80)	.714
Social-emotional	48.99 (28.02)	57.70 (28.43)	.100

*Note.* Standard deviations are presented in parentheses; HOME = Infants and Toddlers Version of the Home Observation for Measurement of the Environment inventory (Caldwell & Bradley, 2003), maximum score = 45; Bayley Scales = Bayley Scales of Infant and Toddler Development (3rd ed.; Bayley, 2006).

<sup>a</sup>  $p$  reported for  $t$  test or chi-square test of treatment versus control groups.



letters in length. According to the documentation included with the intervention, words are selected carefully—although no rubric for selection is provided. Across the different materials, similar words are presented in different fonts and different colors.

Each volume of the intervention includes a DVD, word cards, picture cards, and a picture book. According to the parent guide included with the intervention, infants should watch the DVD at least one time per day and devote at least 15 min with each of the other three materials.

**DVDs.** Each 20-min DVD volume begins with a brief introductory segment in which infants are specifically instructed to attend to the text appearing on-screen. DVDs then follow a routinized format. Lowercase printed words (e.g., *ears*) are shown centered against a solid colored background, taking approximately one third of the visual space. After a 2-s delay, an adult voiceover says a familiar carrier phrase (e.g., “Can you say [target word]”) followed by a child’s voice saying the word that appears on screen. After approximately 2 s, a second child voiceover repeats the onscreen word. Each time the printed word is spoken, a cursor appears beneath the word and travels the length of the word from left to right. The printed word then disappears and is replaced with a scene depicting the meaning of the word (e.g., a child pointing to her ears) accompanied by an adult voiceover describing the scene (e.g., “Katie is pointing to her ears”). For each volume, this pattern (i.e., printed word followed by scene depicting the word) occurs between 60 and 70 times, with each of the target words presented only one time or as many as five times. Each DVD also includes three brief songs or nursery rhymes (e.g., “The Itsy Bitsy Spider”).

**Word cards.** Each volume includes a set of 10–12 flashcards measuring 7.5 by 4.5 in. One word is printed on each side of the card (i.e., 20–24 words total). The words are printed in all lowercase letters in black ink on a colored background; each word is written in a distinct font.

**Picture cards.** Each volume includes five picture cards. Picture cards measure 7.5 by 4.5 in. and have one word written on each side (i.e., 10 words total). Words are printed in lowercase letters in black ink; font and background color vary from word to word. A 1.5-in. tab on the right side can be pulled out to reveal a photograph of the referent against a white background.

**Picture books.** A 16-page picture book is included with each volume of the intervention. Each page includes one target word printed on a large flap. The flap can be lifted to reveal a photograph of the referent. Fonts vary across words; photographs are presented in full color on white backgrounds. There is no additional text other than the word on the page.

## Research Design

Our research was designed to examine the four phases of reading, from pre-alphabetic to consolidated or conventional reading. Although there is substantial agreement among theories in the phases that distinguish its development, Ehri’s four-phase model is the most comprehensive in scope (Ehri, 1994), examining the full developmental period of learning to read. In her model, each phase of reading development is characterized by the predominant type of connection that binds written words to their other identities in memory: (a) *pre-alphabetic*, involving visual and contextual connections (e.g., using word configurations, or word length without any phonological information contributing to the association); (b)

*partial alphabetic*, making connections between some salient letters and sounds (e.g., using the sound values of some letters to form connections between spellings and pronunciation of words); (c) *full alphabetic*, involving complete connections between graphemes and phonemes (e.g., able to form connections between all graphemes in spellings and the phonemes in words, securing the words in memory); and (d) *consolidated alphabetic*, when reading becomes fluent and automatic (e.g., readers can read words as a whole rather than as a sequence of grapheme/phoneme units).

Based on this model, our research design was to examine features of these developmental phases, reasoning that although all children in the sample might not become conventional readers (despite market claims), they would likely show evidence of acquiring some of the skills critical to its development (see Figure 1).

In the present study, therefore, we used multiple methods of assessment including parent reports (language development), as well as 10 eye-tracking tasks throughout the course of 7 months. Eye tracking is a noninvasive methodology permitting high-resolution analyses of eye movement patterns. In addition to overall visual preference, eye tracking allows a more precise analysis of how infants distribute their attention, such as where infants look (i.e., scanning patterns) and how they shift their gaze from one location to another (i.e., saccade latencies; Gredebäck, Johnson, & von Hofsten, 2010). Recognizing that young children might exhibit implicit knowledge prior to explicit demonstrations, these gaze-based measures allowed us to tap visual preferences for orthographic knowledge (Golinkoff, Ma, Song, & Hirsh-Pasek, 2013) and other related reading skills that might otherwise not be recognized in measures that require overt demonstrations (e.g., physical actions or verbal responses).

Over the course of 7 months, we conducted a home visit, arranged for each participant to make four laboratory visits to engage in eye-tracking tasks, and performed monthly assessments of language development. In addition, we held bi-weekly telephone conversations with parents to assess fidelity. Table 2 summarizes these measures, which are described in the following text.

## Baseline Measures

To ensure that our randomization procedures allowed for equivalence between conditions before the intervention, we visited families in their homes and administered the Bayley Scales of Infant and Toddler Development (Bayley, 2006), the HOME Inventory (Caldwell & Bradley, 2003), and a demographic questionnaire.

**Bayley Scales of Infant and Toddler Development.** Infants’ general developmental functioning was assessed using the Bayley Scales of Infant and Toddler Development III (BSID–III), a developmental battery yielding norm-referenced scores (Bayley, 2006). Three domains were examined during the initial home visit. Infants were administered the Cognitive Scale and the Language Scale (including both the Receptive and Expressive Communication subtests) by a trained research assistant, while parents completed the Social–Emotional Scale portion of the caregiver questionnaire. The reported reliability for the BSID–III ranges from .87 to .93.

**Infant/Toddler HOME inventory.** The Infant/Toddler (IT) version of the Home Observation for Measurement of the Environment (HOME; Caldwell & Bradley, 2003) was administered at

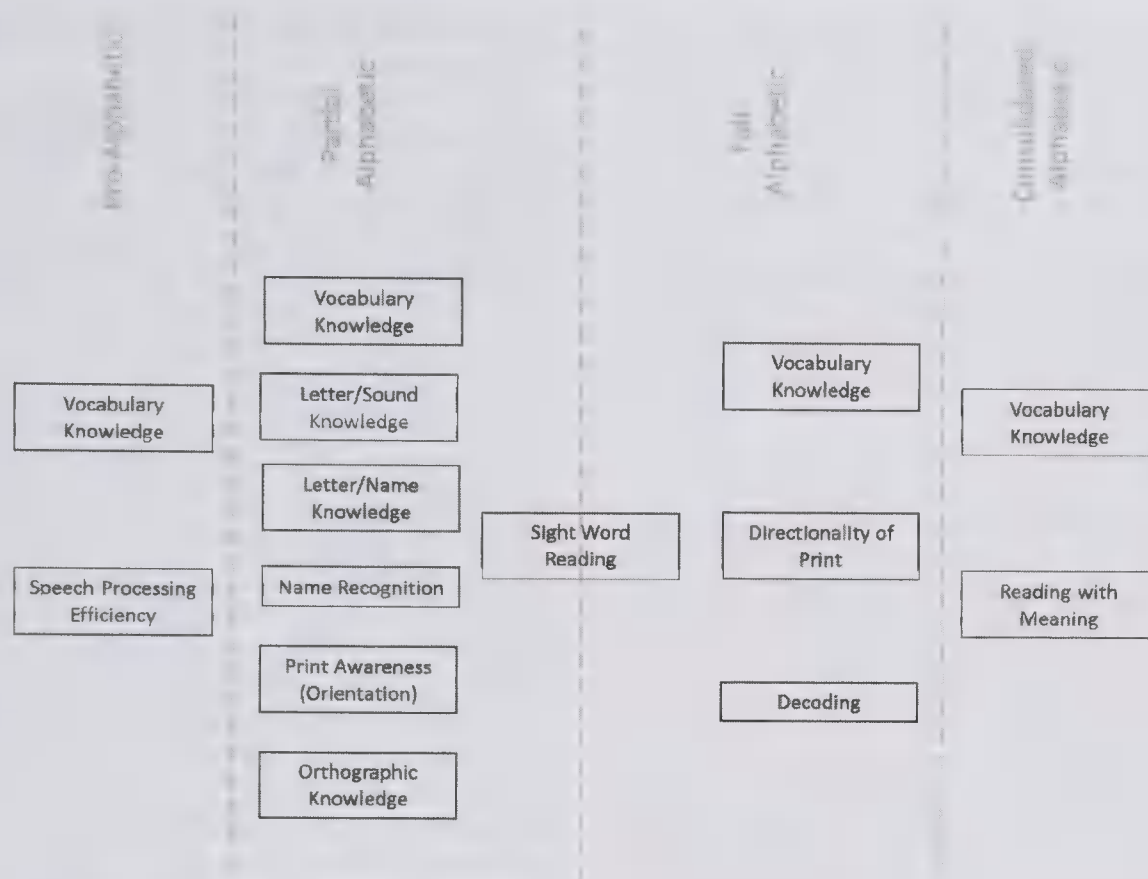


Figure 1. Study outcome measures placed in Ehri's (1994) four-phase model of reading.

the initial home visit. The IT HOME uses both observation and interview to measure the quality and quantity of stimulation and support available to infants in their home environment in six areas: (a) Responsivity, (b) Acceptance, (c) Organization, (d) Learning Materials, (e) Involvement, and (6) Variety. Scores on the sub-

scales were summed (maximum = 45) to yield an overall HOME score.

**Demographic questionnaire.** Parents were asked to complete a 40-item demographic questionnaire. Items included family demographic information (e.g., ethnicity, native language), socioeco-

Table 2

*Developmental Phases of Learning to Read and Their Associated Measures*

Developmental phase <sup>a</sup> /brief description	Measure	Adapted from
Pre-alphabetic		
Expressive vocabulary knowledge	MacArthur CDI	Fenson et al. (1994)
Speed and accuracy of spoken word recognition	Speech processing efficiency	Fernald et al. (2006)
Partial alphabetic		
Receptive vocabulary knowledge	Vocabulary knowledge	Meints et al. (1999); Swingley & Aslin (2000)
Phoneme/grapheme correspondence	Letter-sound knowledge	Letter identification (Woodcock, 1987)
Knowledge of graphemes	Letter-name knowledge	Word attack (Woodcock, 1987)
Identification of written first name	Name recognition	— <sup>b</sup>
Knowledge of standard book format (e.g., upright orientation)	Print awareness (orientation)	DeLoache, Uttal, & Pierroutsakos (2000)
Knowledge of the properties of letters and words	Orthographic knowledge	Cassar & Treiman (1997); Kaefer (2009)
Full alphabetic		
Recognition of previously read words	Sight word reading	Ehri (1994)
Knowledge of words taught in the baby media program	Target vocabulary knowledge	— <sup>b</sup>
Knowledge of standard print format (e.g., right to left, top to bottom)	Directionality of print	Clay (1979)
Translation of text into words	Decoding	Word attack (Woodcock, 1987)
Consolidated alphabetic		
Expressive vocabulary knowledge	MacArthur CDI	Fenson et al. (1994)
Comprehension	Reading with meaning	— <sup>b, c</sup>

Note. MacArthur CDI = MacArthur-Bates Communicative Developmental Inventories.

<sup>a</sup> Adapted from Ehri (1994). <sup>b</sup> Researcher-developed. <sup>c</sup> Adapted from promotional materials for the baby media program.



conomic background (e.g., parent education, family income), family structure (e.g., marital status, number of siblings), and educational activities (e.g., infants' book-reading and baby media experiences).

### Exit Questionnaire

At the end of the study, parents were asked to complete a 16-item questionnaire indicating their beliefs about their children's early literacy behaviors including their ability to write their name, recognize letters, and read.

### Child Assessments

Children's first lab visit occurred approximately 2 weeks after their initial home visit. Follow-up lab visits then occurred after another 3 months, 4 months, and 6 months, or after the treatment group finished their initial viewing of Volumes 2, 3, and 5, respectively. At these visits, we examined the four phases of reading development using a series of eye-tracking tasks.

#### Eye-tracking technology.

**Apparatus.** Eye movements were measured with a T120 eye tracker (Tobii Technology, Falls Church, VA) integrated into a 17-in. thin film transistor (TFT) monitor (Psychology Software Tools, Pittsburgh, PA). This is a remote eye-tracking system that had no contact with the infant. The typical spatial accuracy of this system is approximately 0.5 visual degrees, and the sampling rate is 120 Hz. During tracking, the eye tracker uses infrared diodes to generate reflection patterns on the corneas of the infant's eyes. These reflection patterns, together with other visual information about the infant, are collected by image sensors and used to calculate the three-dimensional position of each eye and gaze point on screen. The TFT monitor employs active matrix technology in which transistors control each pixel on the screen, improving image quality and contrast relative to passive-matrix technologies. The monitor has a display resolution of 96 pixels per inch, ensuring that images are discernible.

This system uses a binocular tracking method, which allows for increased head movements. Head movements typically result in a temporary accuracy error of approximately 0.2 visual degrees. In the case of particularly fast head movements (i.e., over 25 cm/s), there is a 300-ms recovery period to full tracking ability. An embedded camera is also used to record infants' reactions.

**General procedure.** Infants sat in a high chair or on their parent's lap approximately 60 cm from the monitor. Parents wore headphones and blinders to prevent any interference with their infant's looking behavior. Stimuli were displayed on the Tobii monitor and a second monitor facing the experimenter. Tobii Studio Professional 3.0 software was used for stimuli presentation and data processing.

To calibrate gaze, an attention grabber was shown at five points on the screen. A manual calibration procedure was used; accuracy was checked by Tobii Studio software and repeated as necessary. Following calibration, a 2-s attention grabber appeared in the center of the screen prior the beginning of each eye-tracking task. During the task, the experimenter monitored infants' eye movements and behaviors using the live viewer. If infants became distracted during the video, the experimenter made a noise (e.g., snapping or shaking keys) behind the monitor to re-orient their

attention toward the video. Total duration of each eye-tracking task was approximately 5 min.

Each visit took approximately 45–60 min, including both familiarization and testing. In the case when we used both interactive measures, such as the name recognition task and an eye-tracking task, infants would be given a break between tasks. When there were two eye-tracking tasks in a single visit, tasks would be presented consecutively (i.e., no breaks; total testing time approximately 10 min). However, breaks were always given if infants were fussy or if parents requested a break.

Measures were presented in a set order across children. Within each task, trial order was randomized across children (except for name recognition, which was counterbalanced.) If a child was noncompliant on a task, the case would be eliminated for that particular measure (approximately 1% of the time). In the case when a child was entirely noncompliant, the visit would be rescheduled. This only occurred twice (with the same family) over the course of the study.

Children received a small gift, such as a book or toy, at the end of each visit. Parents were compensated \$100 for participation in the study, \$50 at the beginning and \$50 at its completion. In addition, parents received travel expenses (i.e., \$0.56/per mile and parking) and were allowed to keep the baby media product at the conclusion of the study.

**General data processing.** Eye movement data were extracted using Tobii Studio Professional 3.0 software. Fixations were defined as any gaze coordinates lasting at least 60 ms and were identified using the Tobii Studio fixation filter. Adjacent gazes (i.e., gazes within a 0.5° radius, lasting less than 75 ms) were merged into a single fixation.

#### Assessments at the pre-alphabetic phase.

**Expressive vocabulary knowledge.** Infants' expressive vocabulary knowledge was assessed using the short forms of the MacArthur–Bates Communicative Development Inventories (Fenson, Pethick, Renda, & Cox, 2000). Parents of infants ages 16 months and older were asked to indicate which of the words on the Level II form their child produced. Parents of infants younger than 16 months of age completed the Level I form. To facilitate comparison across forms, we computed percentile scores for each child based on his or her age and gender and limited our analysis to expressive vocabulary.

**Speech processing efficiency.** Previous research has indicated that speed of word recognition during infancy is positively predictive of long-term lexical and grammatical development (Fernald, Perfors, & Marchman, 2006). To examine infant's speech processing efficiency, we had the infants view 12 pairs of referents (i.e., target and foil) on the eye-tracking monitor. Referents were selected from published lexical norms (Dale & Fenson, 1996) as familiar to most infants in the age range of our sample. For each trial, a pair of 5 × 5 in. photographs appeared. After a pre-message baseline of 2,150 ms, a voiceover provided a directive (e.g., "Look at the car!"). Photographs then remained on the screen for an additional 2,750 ms. This procedure was then repeated for the remaining trials. Left–right orientation was counterbalanced and presented in a set order; trial order was randomized across infants. Rectangular areas of interest (AOIs) were drawn around each referent. AOIs were kept the same size (461 × 457 pixels) for all photographs for consistency across trials. Fixation duration for each AOI during



the baseline phases and latency to first fixation on the AOIs during the test phase were exported for analysis.

#### **Assessments at the partial alphabetic phase.**

**Receptive vocabulary knowledge.** To assess infants' receptive knowledge of vocabulary introduced in the baby media program, we created an eight-item measure modeled after Meints, Plunkett, and Harris (1999). For this task, infants viewed pairs of referents (i.e., target and foil) on the eye-tracking monitor. All referents were featured in the first two volumes of the baby media program. For each trial, a pair of  $5 \times 5$  in. photographs appeared. Image resolution averaged  $466 \times 520$  pixels. To avoid biases due to color preference, target/nontarget pairs were also broadly matched for color (e.g., a shoe and a hat that were both red).

After a pre-message baseline of 2,150 ms, a voiceover provided a directive (e.g., "Look at the ear!"). Photographs remained on the screen for an additional 2,750 ms. This procedure was then repeated for the remaining trials. Although target words are often presented multiple times in looking-while-listening paradigms (e.g., Fernald, Zangl, Portillo, & Marchman, 2008), we opted for using a single trial per word to potentially avoid task fatigue.

Left-right orientation was counterbalanced and presented in a set order; trial order was randomized across infants. Rectangular AOIs were drawn around each referent; sizes were consistent ( $461 \times 457$  pixels) across trials. Fixation duration for each AOI during the baseline and test phases was exported for analysis.

**Letter-sound knowledge.** To assess infants' understanding of grapheme-sound correspondences, we created a six-item preferential looking task modeled after the Letter-Word Identification subtest of the Woodcock Johnson III (Woodcock, McGrew, & Mather, 2001). For each trial, infants viewed a pair of lowercase letters. After a 2,150-ms pre-message baseline, a voiceover presented a directive (e.g., "Look at the /b/!"). The procedure was then repeated for the remaining trials. Left-right orientation was counterbalanced and presented in a set order; trial order was randomized across infants. Rectangular AOIs were drawn around each referent; sizes were consistent ( $326 \times 312$  pixels) across trials. Fixation duration was then exported for each AOI during the baseline and test phases.

**Letter-name knowledge.** Infants' knowledge of grapheme-name correspondences was assessed through a six-item preferential looking task adapted from the Letter-Word Identification subtest of the Woodcock Johnson III (Woodcock et al., 2001). For each trial, infants viewed a pair of lowercase letters. After a 2,150-ms pre-message baseline, a voiceover presented a directive (e.g., "Look at the t!"). The procedure was then repeated for the remaining trials. Left-right orientation was counterbalanced and presented in a set order; trial order was randomized across infants. After drawing rectangular AOIs of  $326 \times 312$  pixels around the referents, we exported fixation duration for each AOI during the baseline and test phases.

**Name recognition.** Personal names are one of the first written words recognized by young children (Treiman, Cohen, Mulqueeny, Kessler, & Schechtman, 2007). To examine infants' written name recognition, we created a four-item receptive task. For each trial, infants were shown two identical toys (i.e., two green cars or two yellow boats). One toy was labeled with the infant's first name in 20-point font against a white background. The other toy was labeled with a pseudo-word matched in character length with the infant's name (e.g., Nathan vs. Gombie). Infants were

shown both toys and asked to select the one bearing their name (e.g., "Get Nathan's car!"). If infants made a selection, they advanced to the next trial. If infants failed to make a selection or selected both toys, the trial was repeated. This procedure was repeated for a total of two car trials and two boat trials (counterbalanced for toy order and left/right placement). Trials were scored dichotomously (i.e., correct or incorrect), summed to yield an overall score, and converted into a proportion score.

**Print awareness (orientation).** Research suggests that an understanding of the canonical orientation of books and print may be one of the earliest-emerging domains of print awareness (DeLoache, Uttal, & Pierrousakos, 2000). To examine infants' understanding of book and print orientation, we created a six-item preferential-looking task. For half of the trials, infants viewed pairs of book covers (i.e., upright and inverted); for the remaining trials, they viewed pairs of pseudo-words (i.e., upright and inverted). For each trial, infants were oriented to the screen by an attention grabber, and then a pair of images appeared. The trial lasted 10 s; there were no oral prompts. This procedure was then repeated for the remaining trials. Left-right orientation was counterbalanced and presented in a set order; trial order was randomized across infants. For book trials, AOIs of  $450 \times 503$  pixels were drawn around each cover; for word trials, AOIs of  $375 \times 152$  pixels were drawn over each word. Total fixation duration to each AOI was then exported.

**Orthographic knowledge.** Recent work (Kaefer, 2012) suggests that young children's understanding of orthographic conventions may emerge earlier than previously reported. We examined infants' intuitive orthographic knowledge through a nine-item preferential-looking task. In three mirror image trials, we paired a pseudo-word that obeyed English orthographic conventions with its mirror image. In six illegal character trials, we paired orthographically legal pseudo-words (e.g., *pobe*) with orthographically illegal versions (i.e., *p#be*). For each trial, infants were oriented to the screen by an attention grabber, followed by a pair of words. Trials lasted 10 s; there were no oral prompts. Left-right orientation was counterbalanced and presented in a set order; trial order was randomized across infants. Rectangular AOIs were drawn over each word ( $438 \times 159$  pixels), as well as over the individual orthographically illegal characters ( $106 \times 159$  pixels). We then exported total fixation duration to each AOI.

**Sight word reading.** We examined infants' ability to represent familiar sight words in memory (Ehri & Robbins, 1992) through a six-item preferential-looking assessment. For each trial, infants viewed a pair of lowercase words (i.e., target and foil) that were featured in the baby media program. Following a 2,150-ms pre-message baseline, infants were presented with a directive (e.g., "Look at baby!"). Word pairs remained on screen for an additional 2,750 ms. The procedure was then repeated for the remaining trials. Left-right orientation was counterbalanced and presented in a set order; trials were presented in random order across participants. Rectangular  $435 \times 193$  pixel AOIs were drawn around each word, and fixation duration to each AOI during the baseline and test phases was exported.

#### **Assessments at the full alphabetic phase.**

**Target vocabulary knowledge.** Previous research has suggested that young children may acquire little oral vocabulary knowledge from viewing infant-directed DVDs. To assess infants' expressive knowledge of words introduced in the baby media



program, we created a 117-item checklist modeled after the MacArthur–Bates Communicative Development Inventories (Fenson et al., 2007). The checklist consisted of all single words highlighted in at least one volume of the baby media DVD and included in the word cards. Parents were instructed to indicate all words (including their morphological inflections) currently said by their infants. Each checklist was summed to yield an overall score and converted into a proportion score.

**Directionality of print.** Understanding the directionality of text is a key element of young children's developing concepts of print (Justice & Ezell, 2001). In a nine-item task, we assessed infants' understanding of left-to-right directionality of words (six trials) and top-to-bottom directionality of text (three trials). For word trials, infants viewed a single 27- to 30-character word (i.e., directionality) or string of symbols (i.e., no directionality). If infants understood the directionality of text, we would expect them to demonstrate a preference for the AOI for real words and no preference for any of the AOIs for meaningless strings of symbols. For text trials, they viewed several simple sentences taken from commercially available children's books. For each trial, infants were first oriented to the screen by an attention grabber; the trial then lasted 10 s. There were no oral prompts. Trial order was randomized across infants. For word trials, the width of the monitor screen was divided into thirds, and a rectangular AOI ( $318 \times 151$  pixels) was drawn across each third of the text. For sentence trials, the screen was divided into quadrants, and AOIs ( $481 \times 360$  pixels) were drawn to cover each quadrant. Tobii Studio Professional 3.0 software was then used to export the location of first look for each trial.

**Decoding.** Conventional literacy is frequently characterized as the product of code-based skills and comprehension (Gough & Tunmer, 1986). To investigate infants' decoding abilities, we constructed a six-item assessment modeled after the Word Attack subtest of the Woodcock Johnson III (Woodcock et al., 2001). For each trial, infants viewed a pair of words that were not included in the baby media program (i.e., target and foil). Following a 2,150-ms pre-message baseline, infants were presented with a directive (e.g., "Look at cheese!"). The word pair remained on screen for an additional 2,750 ms. This procedure was then repeated for the remaining trials. Left–right orientation was counterbalanced and presented in a set order; trials were presented in random order across participants. Rectangular  $435 \times 193$  pixel AOIs were drawn around each pseudo-word, and fixation duration to each AOI during the baseline and test phases was exported.

#### **Assessments at the consolidated alphabetic phase.**

**Expressive vocabulary knowledge.** Expressive vocabulary knowledge was assessed using the short form of the MacArthur–Bates Communicative Development Inventories (Fenson et al., 2000). At the final visit, all infants were age 16 months or older; therefore, all parents were asked to complete the Level II form. Percentile scores were calculated based on infants' age and gender.

**Reading with meaning.** To examine infants' ability to comprehend simple written phrases, we created a six-item task modeled after the promotional materials for the baby media program. All phrases were featured in at least one volume of the program DVD and word cards.

To ensure that infants understood the task, we first administered two training trials. The research assistant held up a  $5.5 \times 8.5$  in. white card printed with a target phrase in 72-point lowercase text.

While running her finger left to right under the text, she read the depicted phrase aloud (e.g., "It says shake your head!"), orally repeated the target action (e.g., "Shake your head!"), and performed the action. This procedure was repeated until the infant also completed the action.

Following the two training trials, infants completed six test trials (administered in a randomized order). For each trial, the research assistant held up a card and, while running her finger left to right under the text, provided a directive (e.g., "Do this one!"). If infants responded, they were given neutral feedback and moved on to the next trial. If infants failed to respond, the trial was repeated (up to a total of three repetitions). This procedure was then repeated for the remaining test trials. Trials were scored dichotomously (i.e., correct or incorrect) online. Additionally, 20% of video recordings were randomly selected to be independently coded by a second research assistant. Interrater agreement was 95.83%.

## **Procedure**

Following baseline procedures, two trained research assistants visited treatment families in their homes. They introduced the intervention procedure, adapted from the directions provided by the baby media program, modeled its use, and answered any questions. The same researchers visited all families to ensure consistency of instruction.

During the home visit, parents were provided with the first volume of intervention materials. The research assistant reviewed the instructions, suggesting that parents show their infants the first DVD two times per day for 30 days (i.e., 20 hr of exposure). They were encouraged to watch the DVD with their babies and point out words on the screen whenever possible. Parents were also instructed to engage with their infants while using each of the other intervention materials (i.e., word cards, picture cards, and picture book) for 15 min per day (i.e., 7.5 hr of exposure per component) and to feel free to break the interactions up across the day if they found their child was unwilling to attend for the full time. Repetitions of intervention materials were encouraged but not required. Finally, parents were provided with tips for supporting reading activities every day (e.g., playing matching games, pointing out rhyming words). Approximately 30 days after the home visit, families were mailed the second volume of intervention materials, along with a new set of instructions. Following the program's guidelines, parents were instructed to use the second volume once per day for 60 days and to follow the previous protocol's recommendation with each of the other materials. Additionally, parents were asked to show their children the previous DVD, as well as use each of the previous materials, once a week.

By following this protocol designed by the program developers, over the course of the study, infants would receive 70 hr of DVD training and 45 hr of interacting with each of the other materials (i.e., word cards, picture cards, and picture books). Together, they would be exposed to 117 words in multiple formats.

**Fidelity of implementation.** We developed a four-item checklist to capture the degree to which treatment families adhered to the instructions for implementing the baby media program. Using the program's parent guide, we identified key features of the program to include on the checklist: (a) infant watched the DVD, (b) infant used the word cards, (c) infant used the picture cards, and (d) infant read the picture book.

Over the course of the study, a trained research assistant called treatment families on a bi-weekly basis. Parents were asked about their infant's use of the baby media program on the previous day. Checklist items were adapted to specifically target the current volume of the program. Each feature was scored dichotomously as 1 (i.e., completed) or 0 (i.e., not completed). Implementation varied across the four components. Families were mostly likely to use the DVD ( $M = 64.38\%$ ,  $SD = 29.56$ ), followed by the picture book ( $M = 57.92\%$ ,  $SD = 28.99$ ), and picture cards ( $M = 54.48\%$ ,  $SD = 27.64$ ). They were least likely to use the word cards ( $M = 26.90\%$ ,  $SD = 30.07$ ).

To examine whether enactment declined over the course of the study, we compared each family's implementation checklists across the five volumes. Fidelity to the full baby media program (all elements together), as well as implementation of the picture book, picture cards, and word cards, remained consistent across volumes (all  $ps > .05$ ). However, parents reported significantly lower enactment of the final DVD compared with prior volumes ( $p = .014$ ).

## Results

In this section, we address the effects of the intervention on babies' reading development. We begin by conducting descriptive analyses to examine the distributional properties of the data and to determine the equivalency of the treatment and control groups prior to further analysis. Subsequently, we use inferential statistics to examine the effects of the program. To test whether infants who used the components of the program more frequently were more likely to display early literacy skills than children who used the program less frequently, we correlated fidelity with each of the outcome measures. There was no evidence that fidelity to the program supported any of the literacy skills (see Table 3). Further, we conducted all analyses with percentage of fidelity as covariate and found it to be nonsignificant for all measures. Therefore, we used one-way analyses of covariance (ANCOVA) with condition as the independent variable, and baseline as covariate when included in the task, to examine each phase of development in learning to read. Because of the age range among the children and the developmental differences between infants at various stages, we included age as a covariate in all nonstandardized analyses. Additionally, when available, we used pretest or baseline scores as covariates.

## Descriptive Analyses

As shown in Table 1, there were no significant group differences in the child demographic data, HOME scores, or Bayley scores at the infants' first visits.

Treatment and control groups also did not differ initially on their previous media experience. Data for our sample are presented alongside national averages in Table 4. In general, children in our sample had less television exposure than the national average, had fewer televisions in their homes, and were less likely to have cable access than the average child (Rideout & Hammel, 2006).

Contrastingly, infants in our sample in both treatment and control groups were more likely to have shared-reading experience than the national average, with 75% of the parents reporting they read daily to their child (see Table 5). Despite having more regular

Table 3

*Correlation Between Fidelity to the Intervention and Child Outcomes*

Outcome measure	<i>r</i>	<i>p</i>
Expressive vocabulary knowledge	.255	.056
Receptive vocabulary knowledge	-.053	.710
Letter-sound knowledge	-.105	.465
Letter-name knowledge	.048	.743
Name recognition	.200	.168
Print awareness-book orientation	.075	.600
Print awareness-text orientation	-.152	.290
Orthographic knowledge-mirror	.102	.482
Orthographic knowledge-illegal character	.108	.456
Sight word reading	-.283	.045
Target vocabulary knowledge	.227	.102
Directionality-words	.039	.786
Directionality-sentences	.170	.233
Decoding	.132	.356
Expressive vocabulary knowledge	.125	.371
Reading with meaning	-.024	.863

*Note.* None of the measures had  $ps$  greater than the Bonferroni-corrected value of .003.

reading sessions than the national average, infants in our sample appeared to be read to for slightly less duration than the average infant, with the most popular response for parents in our sample being 15–30 minutes, just below the national average of 33 minutes per day.

## Pre-Alphabetic and Partial Alphabetic Phases

Means and standard deviations for the pre-alphabetic and partial alphabetic phases of reading are presented in Table 6. As shown, although the means were slightly higher for the treatment group in the pre-alphabetic phase, there were no significant differences between groups on either of these measures. As evidenced by parental report, expressive vocabulary between groups was statistically equivalent.

**Speech processing efficiency.** Similarly, there were no significant differences between groups in children's ability to process speech. If the treatment had facilitated children's processing of oral language, we would have expected children to demonstrate significantly lower time to first fixation than children in the control group. To measure such processing, we first checked to make sure that children did not have a preference for the target or foil picture prior to being prompted to look. In our task, seven target words passed this first criterion: *bottle*, *bucket*, *car*, *dog*, *ear*, *giraffe*, and *tiger*. Second, to ensure that children's latency to fixate was based on actual processing of the verbal prompt, we needed to establish that children knew the words presented in the task. We confirmed this by asking parents to indicate which of the words their children understood and excluded trials for which we had no parent confirmation. Each child's latency to look to the target was then averaged across each of their eligible trials. There was no difference between groups in latency to look at the target object,  $F(1, 75) = 1.77$ ,  $p = .188$ ,  $\eta^2 = .023$ . The age covariate was also nonsignificant,  $F(1, 75) = 1.33$ ,  $p = .258$ ,  $\eta^2 = .017$ . These results suggest that after more than a month's intervention, there was no evidence of significant effects for treatment on children's pre-alphabetic skills in the ability to process speech more efficiently.



Table 4  
*Infants' Television Exposure*

Questionnaire item	Treatment ( <i>n</i> = 61)	Control ( <i>n</i> = 56)	<i>p</i> <sup>a</sup>	National average
No. of working televisions in the home (%)				
0	3.3	6.0	.343	1% <sup>b</sup>
1–3	90.0	91.0		75% <sup>b</sup>
4 or more	6.7	5.4		24% <sup>b</sup>
Home has cable/satellite television	63.3	69.6	.472	80% <sup>b</sup>
Infant has started watching television	56.7	42.9	.137	79% <sup>c</sup>
Infant's average daily television viewing				34 min <sup>c</sup>
None	64.4	60.7	.446	
< 1 hr	25.4	33.9		
1–2 hr	10.2	5.4		
Parent talks with infant while co-viewing				
Never	14.3	20.5	.177	—
Once in a while	16.3	10.3		
Frequently	53.1	38.5		
Almost always	16.3	30.8		
Infant has television in bedroom	3.3	5.5	.577	19% <sup>c</sup>
Infant has favorite TV program/DVD	41.7	28.6	.140	—
Parent attitudes toward educational media <sup>d</sup>	2.67 (0.43)	2.34 (0.61)	.002**	
Mostly helps <sup>b</sup>				38%
Not much effect <sup>b</sup>				22%
Mostly hurts <sup>b</sup>				31%

<sup>a</sup> *p* reported for *t* test or chi-square test of treatment versus control groups. <sup>b</sup> Data reported by Rideout and Hammel (2006) for children ages 6 months–6-years. <sup>c</sup> Data reported by Rideout and Hammel (2006) for children ages 6-months–23-months. <sup>d</sup> Mean value and standard deviation on scale of 1–4; neutral score is 2.5.

\*\* *p* < .01.

At the partial alphabetic phase, we found a similar pattern. Given the number of tasks at this phase, we would expect that the intervention might influence at least some of the initial speech-to-print skills that it presumably emphasizes throughout the program. However, as shown in Table 5, there were no apparent patterns of improvements in these skills, with the treatment group slightly ahead on four of the measures, and the control group slightly ahead on another four of the measures. None were statistically significant, as described in the following text.

**Receptive vocabulary knowledge.** If children learned the vocabulary words presented in the baby media program, we would expect children in the treatment group to recognize more of the target vocabulary words taken from the program than those in the control group. A one-way ANCOVA indicated that there was no

difference between groups in proportion of time looking to the target photographs at baseline,  $F(1, 101) = 0.56$ ,  $p = .458$ ,  $\eta^2 = .005$ . The age covariate was also nonsignificant,  $F(1, 101) = 0.06$ ,  $p = .805$ ,  $\eta^2 = .001$ . There was also no difference between groups in the proportion of time looking to the target photographs after they were prompted,  $F(1, 102) = 1.32$ ,  $p = .254$ ,  $\eta^2 = .013$ . Age and baseline fixations were both nonsignificant covariates—age:  $F(1, 102) = 0.61$ ,  $p = .439$ ,  $\eta^2 = .006$ ; baseline fixations:  $F(1, 102) = 3.35$ ,  $p = .070$ ,  $\eta^2 = .033$ .

**Letter-name and letter-sound knowledge.** If exposure to the treatment enhanced letter knowledge, we would expect children in this condition to demonstrate significantly longer fixations to the target letter (compared with the foil letter) than children in the control group. One-way ANCOVAs with condition as the

Table 5  
*Infants' Shared Reading Experience*

Questionnaire item	Percentage of those in		<i>p</i> <sup>a</sup>	National average <sup>b</sup>
	Treatment group ( <i>n</i> = 61)	Control group ( <i>n</i> = 56)		
Infant has started being read to	98.4	100	.332	94% <sup>c</sup>
Infant is read to daily	71.7	78.6	.321	58% <sup>c</sup>
Infant's average daily reading duration				33 min
Not at all	1.7	0.0	.485	
A few minutes	25.0	39.3		
At least 15 min	55.0	46.4		
More than 15 min	18.3	14.3		
Infant has favorite book	63.3	58.9	.627	n/a

<sup>a</sup> *p* reported for chi-square test of treatment versus control groups. <sup>b</sup> Rideout and Hammel (2006) for children ages 6-months–23-months.

Table 6  
*Descriptive Statistics for Pre-Alphabetic and Partial Alphabetic Measures*

Measure	Treatment <i>M</i> ( <i>SD</i> )	Control <i>M</i> ( <i>SD</i> )	<i>p</i> <sup>a</sup>
Expressive vocabulary knowledge (Percentile score on the MacArthur CDI short form at home visit)	35.81 (30.17)	31.58 (29.33)	.444
Speech processing efficiency (Latency to look to target in seconds)	0.99 (0.68)	0.80 (0.65)	.188
Receptive vocabulary knowledge (Proportion of time spent looking to target)			
Baseline	.57 (.10)	.56 (.11)	.458
Test	.59 (.12)	.55 (.14)	.254
Letter-sound knowledge (Proportion of time spent looking to target)			
Baseline	.46 (.14)	.49 (.14)	.245
Test	.50 (.16)	.50 (.19)	.968
Letter-name knowledge (Proportion of time spent looking to target)			
Baseline	.48 (.15)	.51 (.15)	.293
Test	.48 (.19)	.44 (.22)	.200
Name recognition (Proportion of trials correct)	.53 (.19)	.59 (.21)	.216
Print awareness (Orientation; proportion of times spent looking to target)			
Books	.54 (.14)	.56 (.12)	.645
Text	.47 (.24)	.47 (.27)	.837
Orthographic knowledge			
Proportion of time spent looking to target words)			
Mirror	.53 (.23)	.52 (.23)	.573
Illegal character	.38 (.15)	.42 (.17)	.203
Looking time to illegal character (in seconds)	0.68 (0.81)	0.46 (0.38)	.071
Sight word reading (Proportion of time spent looking to target)			
Baseline	.49 (.17)	.51 (.17)	.666
Test	.52 (.23)	.50 (.22)	.413

Note. MacArthur CDI short form = MacArthur-Bates Communicative Development Inventories-short form (Fenson et al. (2004).

<sup>a</sup> *p* reported for analysis of covariance comparison of condition.

independent variable and age as the covariate indicated that there was no difference between groups in proportion of time looking to the target letter at baseline for letter sounds,  $F(1, 100) = 1.37$ ,  $p = .245$ ,  $\eta^2 = .014$  or letter names,  $F(1, 100) = 1.12$ ,  $p = .293$ ,  $\eta^2 = .011$ . The age covariate was also nonsignificant in both tests—sounds:  $F(1, 100) = 0.02$ ,  $p = .891$ ,  $\eta^2 < .001$ ; names:  $F(1, 100) = 0.55$ ,  $p = .462$ ,  $\eta^2 = .006$ . This assessment demonstrated that children had no visual preference for one of the letters over the other prior to being prompted where to look.

One-way ANCOVAs with age and baseline looking as covariates indicated that there was also no difference between groups in proportion of time looking at the target letter at test (after the prompt) when the letter sound was prompted,  $F(1, 99) = 0.002$ ,  $p = .968$ ,  $\eta^2 < .001$ . Both covariates were nonsignificant as well—age:  $F(1, 99) = 0.006$ ,  $p = .937$ ,  $\eta^2 < .001$ ; baseline:  $F(1, 99) = 0.63$ ,  $p = .430$ ,  $\eta^2 = .006$ . Additionally, there was no difference between groups in proportion of time looking at the target letter when the letter name was prompted,  $F(1, 95) = 1.67$ ,  $p = .200$ ,  $\eta^2 = .015$ . Age and baseline looking were both significant covariates—age:  $F(1, 95) = 4.14$ ,  $p = .045$ ,  $\eta^2 = .038$ ; baseline:  $F(1, 95) = 8.60$ ,  $p = .004$ ,  $\eta^2 = .078$ . Both groups' proportion looking to the target was near chance (.50) for letter names and sounds.

**Name recognition.** To test whether experience with the intervention improved children's ability to recognize their written names, we analyzed children's performance on four trials in which they were asked to select the toy with their name printed on it. Children received a proportion score for the number of trials in which they chose the correct object out of four. Both groups scored near chance (.50). An ANCOVA with age as the covariate indicated that there was no difference between the treatment and

control groups,  $F(1, 98) = 1.55$ ,  $p = .216$ ,  $\eta^2 = .015$ . Age was also nonsignificant,  $F(1, 98) = 2.23$ ,  $p = .139$ ,  $\eta^2 = .022$ .

**Print awareness.** If experience with the intervention supported children's understanding of print orientation, we would expect children in the treatment condition to look longer at upright book covers and words than children in the control group. However, there was no group difference in proportion of looking to the upright book covers (vs. the inverted ones), controlling for age,  $F(1, 95) = 0.21$ ,  $p = .645$ ,  $\eta^2 = .002$ . Age was a nonsignificant covariate,  $F(1, 95) = 0.045$ ,  $p = .833$ ,  $\eta^2 = .001$ . There was also no difference in proportion of looking to the upright words, controlling for age,  $F(1, 90) = 0.04$ ,  $p = .837$ ,  $\eta^2 = .001$ . Age was again nonsignificant,  $F(1, 90) = 1.12$ ,  $p = .294$ ,  $\eta^2 = .012$ . Both groups spent approximately half the time looking at the target and half the time looking at the foil on both tasks.

**Orthographic knowledge.** Similarly, we would expect children in the intervention to begin to recognize what was or was not a "word." We compared children's proportion looking to two different types of standard and nonstandard pseudo-words. First, we looked at children's preference for standard versus mirror-image-reversed copies of the same word. An ANCOVA with age as the covariate indicated that there was no group difference in proportion looking to the standard word over the reversed word,  $F(1, 93) = 0.32$ ,  $p = .573$ ,  $\eta^2 = .003$ . Age was nonsignificant,  $F(1, 93) = 0.07$ ,  $p = .794$ ,  $\eta^2 = .001$ . Next, we investigated children's preference for standard words versus those with illegal characters inserted (such as # or \$). Again, there were no group differences, controlling for age,  $F(1, 94) = 1.64$ ,  $p = .203$ ,  $\eta^2 = .017$ . Age was nonsignificant,  $F(1, 94) = 0.64$ ,  $p = .426$ ,  $\eta^2 = .007$ . However, we did note that children in the treatment group spent marginally more time (in raw seconds) fixated on the indi-



vidual illegal character than children in the control group, controlling for age,  $F(1, 98) = 3.32, p = .071, \eta^2 = .033$ , indicating some recognition of irregularity. Age was nonsignificant,  $F(1, 99) = 0.01, p = .937, \eta^2 < .001$ .

**Sight word reading.** Finally, if children learned word–text mappings from exposure to the intervention, we would expect to children in the treatment group to demonstrate significantly longer fixations to target words than foil words when prompted. An ANCOVA with age as the covariate indicated that there was no group difference in proportion of looking to the target (vs. the foil) prior to prompting,  $F(1, 100) = 0.19, p = .666, \eta^2 = .002$ . Both groups spent about half the time looking to each side of the screen. These results indicated that children did not prefer to look at one picture over the other prior to being prompted where to focus.

Baseline looking was significant,  $F(1, 98) = 6.14, p = .015, \eta^2 = .059$ . However, a one-way ANCOVA with age and baseline looking as covariates reported that there were no group differences in proportion of looking to the target after prompting,  $F(1, 98) = 0.68, p = .413, \eta^2 = .007$ . Age was a nonsignificant covariate,  $F(1, 98) = 0.00, p = .959, \eta^2 < .001$ .

In sum, these results showed no evidence of positive effects of the intervention on children's pre-alphabetic or partial alphabetic phases of early literacy development. Using multiple measures designed to tap many different aspects of early development, we found no discernable significant differences between groups on word learning or the skills associated with reading development.

### Full Alphabetic and Consolidated Alphabetic Phases

Although one might assume that the latter phases of reading are predicated on improvements in the earlier phases, and not likely to show evidence of impact on children's reading development, we believed it was prudent to examine the skills associated with conventional reading for several reasons. First, there was a rea-

sonable expectation that reading development may not represent a process where one skill is prerequisite for movement to the next phase (Ehri & Roberts, 2006). None of the developmental theories of reading are so rigid. Second, although the instructional design of the program is based on an analytic phonics approach, it focuses mostly on sight word reading and associative learning with words and word meaning connections; and third, claims made by these programs argue for conventional and fluent reading. Therefore, in the final months, we examined the more conventional skills associated with the simple view (e.g., decoding and comprehension). Table 7 presents the means and standard deviations of measures that are representative of the transition to the consolidated phase and conventional reading.

**Directionality of print.** Conceivably, the intervention should help children understand concepts of print, particularly the directionality of text. Therefore, we would expect children in the treatment group to direct their gaze toward the beginning of words and sentences significantly more frequently than children in the control group. We counted the number of trials on which each child's first look was to the upper left portion of the text. Because the data were very heavily skewed (most children did not look left), we opted to use a Kruskal–Wallis test instead of an ANCOVA. The Kruskal–Wallis test is the nonparametric alternative to the analysis of variance that is used when data are nonnormal (Rosner, 2011). It does not allow for the addition of a covariate, so we first checked for a correlation between age and the number of times children's first look was on the left portion of text. There was no correlation between age and number of left looks for word slides or sentence slides. The Kruskal–Wallis test indicated that there were no group differences in the number of left looks for words ( $p = .652$ ) or sentences ( $p = .838$ ).

**Decoding.** If children learned word–text mappings from *Your Baby Can Read*, we would expect children in the treatment group

Table 7  
*Descriptive Statistics for Full Alphabetic and Consolidated Alphabetic Measures*

Measure	Treatment <i>M</i> ( <i>SD</i> )	Control <i>M</i> ( <i>SD</i> )	<i>p</i> <sup>a</sup>
Target vocabulary knowledge (Proportion of target words child says)	.58 (.33)	.41 (.32)	<.001***
Directionality of print (First look to correct position)			
Words			
Mean number	.27 (.67)	.24 (.60)	.652
Range (max = 6)	0–4 correct	0–3 correct	
Sentences			
Mean number	.59 (.83)	.57 (.72)	.838
Range (max = 3)	0–3 correct	0–2 correct	
Decoding (Proportion of time spent looking to target)			
Baseline	.50 (.16)	.54 (.21)	.357
Test	.47 (.20)	.49 (.24)	.744
Expressive vocabulary knowledge (Percentile score on the MacArthur CDI short form at final visit)	46.8 (30.15)	40.3 (32.2)	.289
Reading with meaning (No. of behaviors performed)			
Familiar cues	4	0	.064
Novel cues	2	1	.507

*Note.* MacArthur CDI short form = MacArthur–Bates Communicative Development Inventories–short form (Fenson et al. (2004).

<sup>a</sup> *p* reported for analysis of covariance, Kruskal–Wallis, or Fischer's exact comparisons of condition.

\*\*\* *p* < .001, after controlling for age.

to fixate significantly more on the target words than children in the control group. A one-way ANCOVA, controlling for age, indicated no difference in children's preference for the target over the foil in the baseline phase (prior to being prompted where to look),  $F(1, 100) = 0.86, p = .357, \eta^2 = .008$ . Age was also nonsignificant,  $F(1, 100) = 0.17, p = .679, \eta^2 = .002$ . This established that children had no visual preference for one object over the other. Baseline looking proportion was then entered along with age as a covariate in a one-way ANCOVA to test for group differences in looking to the target during the test phase (after the prompt was given). There was no group difference,  $F(1, 97) = 0.11, p = .744, \eta^2 = .001$ . Age was a nonsignificant covariate,  $F(1, 97) = 0.41, p = .523, \eta^2 = .004$ . Baseline looking was significant,  $F(1, 97) = 4.70, p = .033, \eta^2 = .046$ . This indicated that children's visual preference in the baseline was the best predictor of their looking during test.

**Program vocabulary.** Toward the end of the study, we asked parents in both groups to identify target words that their child could say, using a similar format as the MacArthur–Bates Communicative Development Inventories, only with target words directly from the intervention program. Comparing treatment and control groups, there was a significant main effect of condition on children's expressive knowledge of words introduced in the program,  $F(1, 102) = 5.99, p = .016, \eta^2 = .055$ , after controlling for age. According to parent reports, children in the treatment group knew an average of 58% of the target words ( $SD = 33\%$ ), and the control group knew only 41% ( $SD = 32\%$ ). Age was a significant covariate,  $F(1, 102) = 39.46, p < .001, \eta^2 = .279$ .

**Vocabulary knowledge.** At the same time, we administered the MacArthur–Bates Communicative Development Inventories to both groups in this final phase of the study. Because the percentile ranking is based partially on children's age, age was not used as a covariate in the analysis. Although children in the treatment group were reported to know more words as indicated by the higher mean percentile ranking ( $M = 46.8, SD = 30.15$ ) than children in the control group ( $M = 40.3, SD = 32.2$ ), this difference was nonsignificant,  $F(1, 104) = 1.14, p = .289, \eta^2 = .011$ . Therefore, if children in the treatment group were using more expressive vocabulary than those in the control group, it was likely due to their repeating the words that they heard and saw in the program and not due to a significant increase in vocabulary knowledge at large.

**Reading with meaning.** In our final analysis, we examined children's ability to read with meaning using written cue cards. Three of these cue cards contained exact phrases and actions taught in the program (e.g., "Clap your hands"), and the other three cards contained words used in the program but combined in novel ways (e.g., "Touch your face"). Children's responses were videotaped and coded by the experimenter for whether the child produced the requested action.

There were a total of four positive responses to the familiar cues, all of which were produced by the treatment group. We performed a Fischer's exact test to determine whether the proportion of responses in the treatment group was significantly greater than the null performance of the control group. We used a one-tailed test due to the nature of the directional hypothesis (i.e., babies' being able to read), suspecting that it would be rare and that the opposite (i.e., babies' failing to respond) might be more common. In other words, we did not expect an extreme contingency table indicating common positive (reading) responses. The one-tailed Fischer's

exact significance was  $p = .064$ , indicating that there was a nonsignificant association between positive responses to the program cue cards and condition. For the novel cues, there were only a total of three positive responses, two infants in the treatment group and the other in the control group. The one-tailed Fischer's exact significance was  $p = .746$ , indicating no association between positive responses and treatment group.

We then asked each of the children who had performed at least one of the reading behaviors to return to the lab after 6 months; five of the seven children returned for a visit. We repeated the same set of procedures and prompts as in the initial reading with meaning task. None of the children successfully completed any of the behavioral responses on the written cue cards.

Finally, we examined the exit questionnaire responses of parents of children who had appeared to "read" cards on the task versus the rest of the sample on two exit items: "My child has started to learn to read," and "My child knows how to read." Kruskal–Wallis tests indicated that these parents had given their children higher ratings on both items,  $ps = .021$  and  $.005$ , respectively.

In sum, following the use of the intervention over a 7-month period, there was no evidence to indicate that babies in the treatment group could read or attend to words and texts any differently than children in the control group. Although parents in the treatment group reportedly indicated that children knew significantly more target words in the program than those in the control group, these gains were not evident in the standardized vocabulary measure. Finally, those children who appeared to read and respond to written language cues seen on the DVDs were not able to identify words or phrases after the intervention was completed, despite the fact that their parents believed that they were beginning to read.

## Discussion

The last decade has seen the explosion of baby media targeted to promoting infants' development (Rideout & Hamel, 2006). Among the best-selling products are those that claim to teach babies to read. In their claims, program developers have argued that by accelerating the reading process, their products can enable young children to advance more rapidly in school, reading complex texts which would otherwise be taught in later grades. These developers implicitly make the case that by gaining more knowledge through text, a child can begin to read earlier and that the earlier a child can begin to read, the more proficient he or she will become in school and beyond.

To our knowledge, this is the first study to fully test these claims. Purposely, we set out to be most scrupulous in our definition of reading, examining outcomes that not only included conventional reading but ones that could tap the earlier precursors of literacy skills. To avoid the limitations of previous studies, we followed the program developers' detailed protocol, trained parents in how to use these products, and conducted ongoing fidelity checks to measure and assess dosage. We used outcome measures that included parent reports and a series of eye-tracking tasks that could more sensitively gauge the subtleties of early orthographic and phonological knowledge. Finally, we conducted what is regarded as the "gold standard" in research design, randomly assigning children and their families to treatment and control groups.

Our results indicated that babies did not learn to read. In total, out of 14 different measures of early reading skills, there were 13



null findings. We saw no evidence for the effects on conventional reading, as program developers had indicated on their promotional websites and testimonials, or on any of the pre-alphabetic or partial alphabetic phases of reading. Even with a greater dosage of treatment than in previous studies, there were no effects of the intervention on children's speech processing efficiency, word learning skills, phonological processing, orthographic knowledge, letter recognition, sight word reading, or reading with meaning.

Nevertheless, based on our exit interviews and parent reports of target word learning (i.e., words seen on the DVD), there was the belief among parents that their babies' *were* learning to read and that their children had benefited from the program in their expressive vocabulary development. Parents in the intervention reported that their infants used more of the targeted words from the program than those in the control group. This did not generalize, however, to the standardized measure of expressive word knowledge. In this case, there were no reported differences between groups.

These results suggest that parents may have interpreted imitation and mimicking as an indicator of word learning. A plethora of studies (e.g., Bandura's classic bobo doll experiment; Bandura, 1965; Neuman, 1991), for example, have shown children's ability to mimic what they see on the screen. This mimicking phenomenon was clearly evident in several of the children's responses to written commands on cue cards. Although four children initially responded to a phrase prompt directly from the program immediately after the intervention, none were successful a few months later. Further, none of the children could respond to words in novel phrases. Consequently, testimonials and videos of reading on many of these websites may be preying on parents' wishes and beliefs for their child's precocity and not on the reality of what constitutes meaningful learning. Although we cannot say with full assurance that infants at this age *cannot* learn printed words, we can confidently say that they *did not* learn printed words from a product of this nature.

Our findings provide further support for experimental studies that have demonstrated a lack of significant effects of baby media on receptive language. In two previous studies, for example, neither Robb et al. (2009) nor DeLoache et al. (2010) reported a significant relationship between DVD viewing and infant receptive vocabulary. These results stand in contrast to those of Vandewater (2011) who reported significant positive effects on receptive vocabulary growth using a similar program. Comparing her results with these previous studies, Vandewater argued that the differences in these findings could potentially be attributed to sample size or a sleeper effect in which effects were found 3 months after the intervention. However, in our study, children were likely to be exposed to a far greater dosage than in Vandewater's study (i.e., children in her study saw the video an average of 14 times, whereas ours were instructed to view the initial DVD at least 60 times, with further redundancy built into all of the program materials and repeated throughout the entire five-volume program) with no evident immediate or sleeper effects. Given that Vandewater's results relied on parental report of children's language gains and not direct assessments with children, such findings might reflect a social desirability bias or wishful thinking on the part of parents. DeLoache et al. (2010), who reported both parent opinion and word learning directly assessed with the child, found that parents who enjoyed the DVD overestimated how many words

their children learned from it. Although there is some research to suggest that young children are capable of learning individual words from screen media sources like video (Allen & Scofield, 2010; Krcmar, 2010; Krcmar et al., 2007), it may be a poor substitute for language experiences with live speakers (Golinkoff & Hirsh-Pasek, 1999).

At the same time, we did not find evidence for suppression of language scores, as reported in surveys and correlational studies. Zimmerman, Christakis, and Meltzoff, (2007), for example, reported that watching baby videos predicted significant decrements in vocabulary size for infants between 8 and 16 months. In contrast, we found no declines in vocabulary for either group throughout our study. Similarly, Linebarger and Vaala (2010) have proposed that the expository content in baby videos (i.e., such as the program intervention in this study), compared with narrative or story-like content, may impair language development due to the volume of information presented in these programs. Once again, we found no evidence of decline in language or any other skill as a result of the intervention. Finally, media researchers have suggested that certain features like hearing verbal labels for objects visually depicted on-screen, verbal and visual emphasis on novel words and their referents, and repetition of verbalizations might support increased language acquisition (Krcmar, 2010; Naigles & Mayeux, 2001). However, even with repeated exposure (i.e., 20 words  $\times$  30 days  $\times$  4 types of media), we found little support for language acquisition or skill development from baby media.

Rather, an alternative hypothesis is that babies are neither helped nor harmed by baby media. Clearly, infants can make sense of some video displays. For example, we could not have conducted our eye-tracking studies without some sustained attention to screen media. Moreover, studies comparing video instruction with no instruction at all have consistently shown that infants do learn some information from videos (Barr & Hayne, 1999; Strouse & Troseth, 2008). Further, there are a number of impressive studies suggesting that older infants and toddlers, given the right type of experiences, can bring their perceptual skills and general knowledge backgrounds to bear on a novel problem-solving tasks presented through screen media (Nielsen, Simcock, & Jenkins, 2008; Troseth, Saylor, & Archer, 2006). Nevertheless, the absence of any significant effects on children's language and skills suggests to us that the developmentalists are most accurate in their recognition of infant capabilities and limitations. Infants do bring a limited set of experiences and little background knowledge of the content and format used to deliver information on screen. As a result, they are in poor position to use screen media as a tool for learning how to read, given that language development and background knowledge are required for reading performance even at the initial level (Neuman, 2006).

Ultimately, therefore, it is about choice. Parents must weigh whether such exposure to media is displacing other activities (Neuman, 1991), such as adult-child language interaction, reading books, play, and joint activity. These are the activities, shown through a large convergence of research (Neuman & Celano, 2012; Snow, Burns, & Griffin, 1998), that have strong empirical support on children's affect, cognitive development, early reading skills, and, in the long run, reading performance.

## References

- Allen, R., & Scofield, J. (2010). Word learning from videos: More evidence from 2-year-olds. *Infant and Child Development*, 19, 649–661. doi:10.1002/icd.712
- American Academy of Pediatrics. (1999). Media education. *Pediatrics*, 104, 341–343. doi:10.1542/peds.104.2.341
- Anderson, D., & Pempek, T. (2005). Television and very young children. *American Behavioral Scientist*, 48, 505–522. doi:10.1177/0002764204271506
- Bandura, A. (1965). Influence of models' reinforcement contingencies on the acquisition of imitative responses. *Journal of Personality and Social Psychology*, 1, 589–595. doi:10.1037/h0022070
- Barr, R., Danziger, C., Hilliard, M., Andolina, C., & Ruskis, J. (2009). Amount, content, and context of infant media exposure: A parental questionnaire and diary analysis. *International Journal of Early Years Education*, 18, 107–122. doi:10.1080/09669760.2010.494431
- Barr, R., & Hayne, H. (1999). Developmental changes in imitation from television during infancy. *Child Development*, 70, 1067–1081. doi:10.1111/1467-8624.00079
- Barr, R., Muentener, P., Garcia, A., Fujimoto, M., & Chávez, V. (2007). The effect of repetition on imitation from television during infancy. *Developmental Psychobiology*, 49, 196–207. doi:10.1002/dev.20208
- Bayley, N. (2006). *Bayley Scales of Infant and Toddler Development* (3rd ed.). San Antonio, TX: PsychCorp/Pearson.
- BrillKids. (2011). *Why teach reading early?* Retrieved from <http://www.brillbaby.com/teaching-baby/reading/why-teach-reading-early.php>
- Caldwell, B. M., & Bradley, R. H. (2003). *Home Observation for Measurement of the Environment: Administration manual*. Tempe: Arizona State University, Family & Human Dynamics Research Institute.
- Cassar, M., & Treiman, R. (1997). The beginnings of orthographic knowledge: Children's knowledge of double letters in words. *Journal of Educational Psychology*, 89, 631–644. doi:10.1037/0022-0663.89.4.631
- Chall, J. (1983). *Stages of reading development*. New York, NY: McGraw Hill.
- Clay, M. (1979). *The early detection of reading difficulties*. Portsmouth, NH: Heinemann.
- Dale, P. S., & Fenson, F. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, 28, 125–127. doi:10.3758/BF03203646
- DeLoache, J. (2004). Becoming symbol-minded. *Trends in Cognitive Science*, 8, 66–70. doi:10.1016/j.tics.2003.12.004
- DeLoache, J., Chiong, C., Sherman, K., Islam, N., Vanderborght, M., Troseth, G., . . . O'Doherty, K. (2010). Do babies learn from baby media? *Psychological Science*, 21, 1570–1574. doi:10.1177/0956797610384145
- DeLoache, J. S., Uttal, D. H., & Pierroutsakos, S. L. (2000). What's up? The emergence of an orientation preference for picture books. *Journal of Cognition and Development*, 1, 81–95. doi:10.1207/S15327647JCD0101N\_9
- Doman, G., & Doman, J. (2010). *How to teach your baby to read*. Wyndmoor, PA: Gentle Revolution Press.
- Ehri, L. C. (1979). Linguistic insight: Threshold of reading acquisition. In T. G. Waller & G. F. MacKinnon (Eds.), *Reading research: Advances in theory and practice* (Vol. 1, pp. 63–111). New York, NY: Academic Press.
- Ehri, L. (1994). Development of the ability to read words: Update. In R. Ruddell, M. R. Ruddell, & H. Singer (Eds.), *Theoretical models and processes of reading* (4th ed., pp. 323–358). Newark, DE: International Reading Association.
- Ehri, L. C., & Robbins, C. (1992). Beginners need some decoding skill to read words by analogy. *Reading Research Quarterly*, 27, 12–26. doi:10.2307/747831
- Ehri, L., & Roberts, T. (2006). The roots of learning to read and write: Acquisition of letters and phonemic awareness. In D. Dickinson & S. B. Neuman (Eds.), *Handbook of early literacy research* (pp. 113–134). New York, NY: Guilford Press.
- Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D., & Pethick, S. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, 59(5), Serial No. 242.
- Fenson, L., Marchman, V., Thal, D., Dale, P., Resnick, J. S., & Bates, E. (2007). *MacArthur–Bates Communicative Development Inventories*. Baltimore, MD: Brookes.
- Fenson, L., Pethick, S., Renda, C., & Cox, J. L. (2000). Short-form versions of the MacArthur Communicative Development Inventories. *Applied Psycholinguistics*, 21, 95–116. doi:10.1017/S0142716400001053
- Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental Psychology*, 42, 98–116. doi:10.1037/0012-1649.42.1.98
- Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In I. Sekerina, E. M. Fernández, & H. Clahsen (Eds.), *Developmental psycholinguistics: On-line methods in children's language processing* (pp. 97–135). Amsterdam, the Netherlands: Benjamins.
- Garrison, M., & Christakis, D. (2005). *A teacher in the living room? Educational media for babies, toddlers, and preschoolers*. Menlo Park, CA: Kaiser Family Foundation.
- Golinkoff, R. M., & Hirsh-Pasek, K. (1999). *How babies talk: The magic and mystery of language in the first three years of life*. New York, NY: Dutton.
- Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: What have we learned? *Perspectives on Psychological Science*, 8, 316–339. doi:10.1177/1745691613484936
- Goodman, Y. (1984). The development of initial literacy. In H. Goelman, A. Oberg, & F. Smith (Eds.), *Awakening to literacy* (pp. 102–109). Exeter, NH: Heinemann.
- Gough, P., & Tunmer, W. (1986). Decoding, reading, and reading disabilities. *Remedial and Special Education*, 7, 6–10. doi:10.1177/074193258600700104
- Gredebäck, G., Johnson, S., & von Hofsten, C. (2010). Eye tracking in infancy research. *Developmental Neuropsychology*, 35, 1–19. doi:10.1080/87565640903325758
- Intellectual Baby. (2009). *Can babies really learn to read?* Retrieved from [http://www.intellbaby.com/teach\\_your\\_baby\\_to\\_read.html](http://www.intellbaby.com/teach_your_baby_to_read.html)
- Justice, L., & Ezell, H. (2001). Word and print awareness in 4-year old children. *Child Language Teaching and Therapy*, 17, 207–225. doi:10.1191/026565901680666527
- Kaefer, T. (2009). *Implicit, eclipsed, but functional: The development of orthographic knowledge in early readers*. (Doctoral dissertation). Available from the ProQuest Dissertations & Theses database. (AAT 3366786)
- Kaefer, T. (2012). What you see is what you get: Learning from the ambient environment. In A. M. Pinkham, T. Kaefer, & S. B. Neuman (Eds.), *Knowledge development in early childhood: Sources of learning and classroom implications* (pp. 3–17). New York, NY: Guilford Press.
- Krcmar, M. (2010). Can social meaningfulness and repeat exposure help infants and toddlers overcome the video deficit? *Media Psychology*, 13, 31–53. doi:10.1080/15213260903562917
- Krcmar, M., Grela, B., & Lin, K. (2007). Can toddlers learn vocabulary from television? An experimental approach. *Media Psychology*, 10, 41–63. doi:10.1080/15213260701300931
- Kuhl, P., Tsao, F., & Liu, H. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction of pho-



- netic learning. *Proceedings of the National Academy of Sciences, USA*, 100, 9096–9101. doi:10.1073/pnas.1532872100
- Linebarger, D., & Vaala, S. (2010). Screen media and language development in infants and toddlers: An ecological perspective. *Developmental Review*, 30, 176–202. doi:10.1016/j.dr.2010.03.006
- Mason, J. (1980). When do children begin to read: An exploration of four year old children's word reading competencies. *Reading Research Quarterly*, 15, 203–227. doi:10.2307/747325
- Masonheimer, P., Drum, P., & Ehri, L. (1984). Does environmental print identification lead children into word reading? *Journal of Reading Behavior*, 16, 257–272. doi:10.1080/10862968409547520
- Meints, K., Plunkett, K., & Harris, P. L. (1999). When does an ostrich become a bird? The role of typicality in early word comprehension. *Developmental Psychology*, 35, 1072–1078. doi:10.1037/0012-1649.35.4.1072
- Naigles, L. R., & Mayeux, L. (2001). Television as incidental language teacher. In D. G. Singer & J. L. Singer (Eds.), *Handbook of children and the media* (pp. 135–152). Thousand Oaks, CA: Sage.
- National Early Literacy Panel. (2008). *Developing early literacy: Report of the National Early Literacy Panel*. Washington, DC: National Institute for Literacy.
- National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. (NIH Pub. No. 00–4754). Washington, DC: National Institute of Child Health and Human Development.
- Neuman, S. B. (1991). *Literacy in the television age*. Norwood, NJ: Ablex.
- Neuman, S. B. (2006). The knowledge gap: Implication for early literacy development. In D. Dickinson & S. B. Neuman (Eds.), *Handbook of early literacy research* (pp. 29–40). New York, NY: Guilford Press.
- Neuman, S. B., & Celano, D. (2012). *Giving our children a fighting chance: Affluence, literacy, and the development of information capital*. New York, NY: Teachers College Press.
- Nielsen, M., Simcock, G., & Jenkins, L. (2008). The effect of social engagement on 24-month-olds' imitation from live and televised models. *Developmental Science*, 11, 722–731. doi:10.1111/j.1467-7687.2008.00722.x
- Rice, M. (1983). The role of television in language acquisition. *Developmental Review*, 3, 211–224. doi:10.1016/0273-2297(83)90030-8
- Rideout, V. J. (2007). *Parents, children and media*. Menlo Park, CA: Kaiser Family Foundation.
- Rideout, V. J., & Hammel, E. (2006). *The media family: Electronic media in the lives of infants, toddlers, preschoolers and their parents*. Menlo Park, CA: Kaiser Family Foundation.
- Robb, M., Richert, R., & Wartella, E. (2009). Just a talking book? Word learning from watching baby videos. *British Journal of Developmental Psychology*, 27, 27–45. doi:10.1348/026151008X320156
- Roseberry, S., Hirsh-Pasek, K., Parish-Morris, J., & Golinkoff, R. (2009). Live action: Can young children learn verbs from video? *Child Development*, 80, 1360–1375. doi:10.1111/j.1467-8624.2009.01338.x
- Rosner, B. (2011). *Fundamentals of biostatistics* (7th ed.). Boston, MA: Cengage Learning.
- Snow, C., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.
- Stahl, S., & Murray, B. (1993). Environmental print, phonemic awareness, letter recognition, and word recognition. In D. Leu & C. Kinzer (Eds.), *Examining central issues in literacy research, theory, and practice* (pp. 227–233). Chicago, IL: National Reading Conference.
- Strouse, G. A., & Troseth, G. L. (2008). "Don't try this at home": Toddlers' imitation of new skills from people on video. *Journal of Experimental Child Psychology*, 101, 262–280. doi:10.1016/j.jecp.2008.05.010
- Strouse, G. A., & Troseth, G. L. (2013). *Supporting toddlers' transfer of word learning from video*. Manuscript submitted for publication.
- Swingle, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, 76, 147–166. doi:10.1016/S0010-0277(00)00081-0
- Teale, W., & Sulzby, E. (1989). Emergent literacy: New perspectives. In D. Strickland & L. M. Morrow (Eds.), *Emerging literacy: Young children learn to read and write* (pp. 1–15). Newark, DE: International Reading Association.
- Titizer, R. (2010). *Your baby can read! Early language development system*. Carlsbad, CA: Your Baby Can.
- Treiman, R., Cohen, J., Mulqueeny, K., Kessler, B., & Schechtman, S. (2007). Young children's knowledge about printed names. *Child Development*, 78, 1458–1471. doi:10.1111/j.1467-8624.2007.01077.x
- Troseth, G. L., Saylor, M. M., & Archer, A. H. (2006). Young children's use of video as a source of socially relevant information. *Child Development*, 77, 786–799. doi:10.1111/j.1467-8624.2006.00903.x
- Troseth, G. L., Strouse, G. A., Verdine, B. N., & Saylor, M. M. (2013). *Responsiveness and relevance: Is social information from video sufficient for word learning?* Unpublished manuscript, Department of Psychology and Human Development, Vanderbilt University, Nashville, TN.
- Vandewater, E. (2011). Infant word learning from commercially available video in the U.S. *Journal of Children and Media*, 5, 248–266. doi:10.1080/17482798.2011.584375
- Woodcock, R. W. (1987). *Woodcock Reading Mastery Tests-Revised: Examiner's manual*. Circle Pines, MN: American Guidance Service.
- Woodcock, R. W., McGrew, K., & Mather, N. (2001). *Woodcock Johnson III Tests of Achievement*. Itasca, IL: Riverside.
- Zimmerman, F., Christakis, D., & Meltzoff, A. (2007). Associations between media viewing and language development in children under age 2 years. *Journal of Pediatrics*, 151, 364–368. doi:10.1016/j.jpeds.2007.04.071

Received July 21, 2013

Revision received October 16, 2013

Accepted November 25, 2013 ■

# Does Cognitive Strategy Training on Word Problems Compensate for Working Memory Capacity in Children With Math Difficulties?

H. Lee Swanson  
University of California, Riverside

Cognitive strategies are important tools for children with math difficulties (MD) in learning to solve word problems. The effectiveness of strategy training, however, depends on working memory capacity (WMC). Thus, children with MD but with relatively higher WMC are more likely to benefit from strategy training, whereas children with lower WMC may have their resources overtaxed. Children in Grade 3 ( $N = 147$ ) were randomly assigned to 1 of 4 conditions: (a) verbal strategies (e.g., underlining question sentence), (b) visual strategies (e.g., correctly placing numbers in diagrams), (c) verbal plus visual strategies, or (d) an untreated control. In line with the predictions, children with MD and higher WMC benefited from verbal or visual strategies relative to those in the control condition on posttest measures of problem solving, calculation, and operation span. In contrast, cognitive strategies decreased problem-solving accuracy in children with low WMC. Thus, improvement in problem solving and related measures, as well as the impairment in learning outcomes, was moderated by WMC.

*Keywords:* math disabilities, strategy training, working memory

The majority of the research on children who experience math difficulties (MD) has focused on processes related to calculation (Andersson, 2010; Geary, 2003, 2010; Gersten et al., 2009; Mazzocco, Devlin, & McKenney, 2008; Stock, Desoete, & Roeyers, 2010; Swanson & Jerman, 2006). More recent studies, however, have focused on children who experience difficulties solving word problems (e.g., Andersson, 2010; Fuchs, Zumeta, et al., 2010; Stock et al., 2010; Swanson, Jerman, & Zheng, 2008). This is an important focus because word problem solving constitutes one of the most critical mediums through which children can learn to select and apply strategies for coping with everyday problems. In addition, recent studies have shown that the cognitive processes involved in calculation difficulties are not the same processes as those involved in problem-solving difficulties (e.g., Fuchs et al., 2008) and therefore call for unique interventions. In addition, some studies have shown that deficits in word problem solving are persistent across the elementary school years even when calcula-

tion and reading skills are at grade level (e.g., Swanson et al., 2008).

Recent intervention studies directed to improve problem-solving accuracy in children with MD have found support for teaching cognitive strategies. Several studies have found that verbal strategy instruction (e.g., Montague, 2008; Montague, Warger, & Morgan, 2000; Xin, 2008), as well as visual-spatial strategies (e.g., Kollhoffel, Eysink, de Jong, & Wilhelm, 2009; van Garderen, 2007), enhance children's math performance relative to control conditions (see Baker, Gersten, & Lee, 2002; Gersten et al., 2009 for reviews). Several well-designed intervention studies (randomized clinical trials) have focused on high-risk samples. For example, Jitendra et al. (1998) used a visual categorization method to cluster arithmetic word problems (e.g., change, compare) that significantly improved problem-solving accuracy compared to the control condition (effect size 0.45). Likewise, in a randomized control group design, Fuchs et al. (2003) taught problem-solving methods to children with MD and found that cognitive strategies (schema-based instructions) improved problem-solving accuracy (effect sizes ranged from 1.16 to 1.18 depending on the transfer measure). Additional successful strategy models have included diagramming (van Garderen, 2007), identification of key words (e.g., Mastropieri, Scruggs, & Shiah, 1997), and meta-cognitive strategies (e.g., Montague, 2008; see Gersten et al., 2009; Xin & Jitendra, 1999, for reviews). These studies strongly suggest that the training of cognitive strategies facilitates problem-solving accuracy in children with MD.

Despite the overall benefits of strategy instruction in remediating word-problem-solving word difficulties, the use of strategies for some children with MD may not always be advantageous. From an aptitude-treatment perspective, not all children with MD may be expected to benefit from strategy training. In this study, I hypothesize that the availability of ample working memory resources is an important precondition in determining whether strat-

---

This article was published Online First February 10, 2014.

This paper was supported by a second-year goal 2 grant funded by the U.S. Department of Education, Special Education: Cognition and Student Learning (USDE R3234A090002), Institute of Education Sciences, awarded to the author. The author is greatly indebted to Cathy Lussier, who served as project director, and to Loren Alberg, Garrett Briney, Kristi Bryant, Olga Jerman, Orheta Rice, Dennis Sisco-Taylor, Catherine Tung, and Kenisha Williams in the data collection and/or task/curriculum development. Special appreciation is given to school administrators Sandra Briney, Jan Gustafson-Corea, and Chip Kling. The report does not necessarily reflect the views of the U.S. Department of Education or the school districts.

Correspondence concerning this article should be addressed to H. Lee Swanson, Department of Educational Psychology, Graduate School of Education, University of California, Riverside, Riverside, CA 92521. E-mail: Lee.Swanson@ucr.edu



egy training will be successful. This is because strategies are resource demanding. As a consequence, children with relatively smaller working memory capacities (WMC) may be easily overtaxed by certain strategies, which may even lead to poor learning outcomes after training. This is because word problem solving is an activity that draws upon WMC to a considerable degree. Because children with MD experience working memory difficulties (e.g., Swanson & Beebe-Frankenberger, 2004), children with low WMC may be unable to effectively benefit from cognitive strategy interventions. In contrast, children with MD who meet a certain threshold of (yet to be determined) WMC would have spare working memory resources to benefit from cognitive strategies. This hypothesis is in line with cognitive load theory (e.g., Sweller, 1988, 2005), whose central tenet is that instruction should be designed in alignment with the learners' cognitive architecture, which consists of a limited-capacity working memory system. Because information has to pass through working memory before it can be consolidated into long-term memory, the limited capacity of working memory can be considered the bottleneck for learning. Thus, individuals with MD but relatively higher WMC are better able than children with lower WMC to utilize cognitive strategies. This is because strategies rely on declarative representations and serial cognitive processes that require large amounts of WMC (e.g., Anderson, 1987), and the utilization of cognitive strategies that have been recently acquired imposes demands on WMC. In the context of this study, I define working memory as a processing resource of limited capacity that is involved in the preservation of information while simultaneously processing the same or other information (e.g., Baddeley & Logie, 1999; Engle, Tuholski, Laughlin, & Conway 1999).

Although the above hypothesis is plausible, there are at least three alternative possibilities to explain the role of WMC and the utilization and training of cognitive strategies to enhance problem-solving accuracy in children with MD. First, individual differences in WMC may not moderate the use of cognitive strategies in children with problem-solving difficulties. Support for this position comes from studies showing that WMC is not predictive of problem-solving accuracy when basic skills, such as reading, are entered into the regression model (e.g., Lee, Ng, Ng, & Lim, 2004; Swanson, Cooney, & Brock, 1993). Further, it could be argued that strategy training is primarily directed at providing additional cues to facilitate the retrieval of computational information via comprehension, and therefore word problem solving does not interact with WMC. Thus, WMC would not moderate the impact of strategy training on problem solving because the impact of strategy use during problem solving is through a route unaffected by WMC. In the second alternative, WMC operates as a general system that subsumes many higher and lower order processes related to word problem solving accuracy. That is, processes related to word problem solving (computation and comprehension) share resources with working memory (e.g., Colom, Abad, Quiroga, Shih, & Flores-Mendoza, 2008; Engle, Cantor, & Carullo, 1992). For example, WMC is necessary in the solving of word problems to allow for (or provide resources for) the translation of numbers and text information into algorithms, as well as for the simultaneous storage of output from previous processing. Thus, WMC predicts problem-solving performance but does not directly interact with cognitive strategies in facilitating problem-solving performance. Thus, there is no direct moderation of strategy effects by WMC.

In the final alternative, cognitive strategies compensate for the excessive processing demands placed on WMC due to the extraneous load of the problem-solving task. For example, solving a word problem (e.g., "15 dolls are for sale, 7 dolls have hats. The dolls are large. How many dolls do not have hats?") involves a variety of mental activities. Children must access prestored information (e.g., 15 dolls), access the appropriate algorithm (15 minus 7), and apply problem-solving processes to control its execution (e.g., ignore the irrelevant information). The multistep nature of word problems that requires the processing of both relevant and irrelevant propositions draws on WMC. Children with relatively low WMC prior to training may be more responsive to cognitive strategies because such strategies help them compensate for working memory limitations. In contrast, children with relatively higher levels of WMC may experience a level of redundancy or unnecessary processing related to strategy training that does not facilitate learning. Thus, this alternative hypothesis predicts that WMC moderates strategy outcomes, but the effects are different than the first hypothesis. The first hypothesis predicts that strategy effects are greatest for children high in WMC, whereas the latter hypothesis predicts that strategy training is more effective for children low in WMC.

This study investigates the role of WMC in strategy training for children with MD, by comparing three cognitive interventions to boost word problem solving performance. Training provided explicit instruction related to (a) verbal strategies that directed children to identify (e.g., via underlining and circling) relevant or key propositions within the problems, (b) visual strategies that required children to place numbers into diagrams, or (c) a strategy condition that combined verbal and visual strategies. Also, because warm-up activities related to calculation have been found to be effective in problem-solving interventions, this component was also included in all strategy training sessions (e.g., Fuchs et al., 2003). In addition, consistent with literature reviews that have identified key components related to treatment effectiveness (Gersten et al., 2009; Xin & Jitendra, 1999), each strategy training session involved explicit practice and feedback related to strategy use and performance. The strategy conditions also directed children's attention to the relevant propositions within each word problem (Mayer & Hegarty, 1996). Additionally, embedded within each lesson were instructions to focus on relevant information for solution accuracy in the context of where there were increasing distractions related to number of irrelevant propositions within the word problems. This is an important component because difficulties in controlled attention have been found to underlie some of the cognitive deficits experienced by children with MD (e.g., Passolunghi, Cornoldi, & De Liberto, 2001; Passolunghi & Siegel, 2001).

In summary, this study addresses the question, What role does working memory capacity (WMC) play in strategy training outcomes for children with MD? Four prediction models based on the aforementioned hypotheses can be applied to strategy training outcomes for children with MD: (a) WMC as a limiting factor, (b) basic skills, (c) general resource, and (d) compensatory. The hypothesis that WMC serves as a limiting factor suggests that children with lower WMC benefit less from strategy conditions than children with relatively higher WMC. Thus, children with MD vary in their responsiveness to strategy instruction, and this is predicated on their WMC. In contrast, the basic skills model suggests that if declarative knowledge is intact (i.e., reading com-



prehension and computational knowledge are in the average range), strategy instruction provides a helpful procedure to solve word problems without making demands on WMC. This model suggests that cognitive instruction provides additional information over control conditions when basic skills (e.g., calculation, reading) are intact. For example, children with MD benefit from strategy instruction because they are less efficient than average achieving children in calculation and general problem solving. Thus, strategy instruction interacts with general math ability and not WMC. In contrast, the general resource model hypothesis predicts that individual differences in WMC are related to solution accuracy regardless of treatment conditions. The resource model predicts that because WM as a general system underlies several problem-solving tasks, WMC has a general effect (nontreatment-specific effect) on problem-solving outcomes. Finally, the compensatory model suggests that WM interacts with treatment outcomes. Cognitive training is viewed as reducing processing demands on children's problem solving and therefore freeing additional resources to solve problems. The compensatory model predicts that children with low WMC are more likely than those with relatively higher WMC to place a greater reliance on strategy conditions.

The first hypothesis predicts an interaction in favor of the high WMC group; the second predicts no significant involvement of WMC in strategy outcomes (no significant main effect or interaction); the third predicts a main effect for WMC but no interaction with strategy conditions; and the fourth predicts a significant WMC by cognitive strategy interaction in favor of children with low WMC.

## Method

### Participants

Participants were 147 third-grade children from public school classrooms in the southwestern United States. The final selection was based on parent approval for participation and achievement scores.<sup>1</sup> Of the 147 children selected, 74 were female and 73 were male. Ethnic representation of the sample was 83 Anglo, 30 Hispanic, 13 African American, 8 Asian, and 13 mixed and/or other (e.g., Anglo and Hispanic, Native American). The mean socioeconomic status (SES) of the sample was primarily low SES to middle SES based on free and reduced lunch participation, parent education, and parent occupation. The schools provided the percentage of children within classroom on a free-lunch program but not for individual participants. Significant differences occurred across classrooms ( $N = 22$ ) in terms of the percentage of free-lunch representation (percentages varied from 2% to 56% of the classes),  $\chi^2(9, N = 22) = 73.62, p < .01$ . However because children were randomly assigned to treatments within classrooms, I assumed that SES was not a contributing factor to the treatment outcomes. Based on the school records, the sample was drawn randomly from classrooms that reflected low middle class to upper middle class.

**Definition of risk for math difficulties (MD).** This study sought to identify children at risk for difficulties in problem-solving performance. There is some consensus among researchers that it is more appropriate to use a cutoff score on achievement to determine risk factors in math rather than a discrepancy between

achievement and IQ. Therefore, this study uses a cutoff score on standardized math achievement tests. Because the majority of children were not diagnosed with specific learning disabilities in math, I utilized the term "at risk for math difficulties" to indicate math difficulty (MD). Because this study's focus was on word problem solving difficulties, I examined children who performed in the lowest 25th percentile on norm-referenced word problem solving math tests over a 2-year period. The 25th percentile cutoff score on standardized achievement measures has been commonly used to identify children at risk (e.g., Fletcher et al., 1989; Siegel & Ryan, 1989). This procedure separated the sample into 59 children with MD (25 female, 34 male) and 88 children without MD (50 female, 38 male). I chose to focus this intervention on children with MD at the third-grade level because this is when word problems are emphasized within the curriculum relative to the early grades.

The cutoff criteria for defining children at risk for MD was a score between the 35th and 90th percentile on measures of fluid intelligence (Raven Colored Progressive Matrices Test; Raven, 1976), reading (Test of Reading Comprehension, Word Identification subtest from the Wide Range Achievement Test (WRAT-3; Wilkinson, 1993), and calculation (subtests from the WRAT-3 and Wechsler Individual Achievement Test; Psychological Corporation, 1992), in addition to a composite score at or below the 25th percentile (below a standard score of 90 or scale score of 8) on standardized word problem solving math tests. Children were considered at risk if they performed at or below the 25th percentile on two of the problem-solving subtests. The story problem subtests were taken from the Test of Math Ability (TOMA; Brown, Cronin, & McEntire, 1994) and KeyMath (Connolly, 1998). Table 1 shows the means and standard deviations for children with MD and average achievers. As shown in Table 1, performance on standardized measures of word problem solving accuracy for the MD sample was at or below the 25th percentile (scale score at or below 8, standard score below 90), whereas the MD sample's norm-

<sup>1</sup> The sample was selected from two charter schools as part of a large Board of Cooperative Educational Services (BOCES). The charter schools serve a large number of children with learning disabilities as well as a major clinical site for the special education local plan area (SELPA) that has a higher than average population of children with special needs (20% of the school population). This allowed selection of participants within each classroom to be randomly assigned to each treatment condition. Although 210 children participated in the study, final selection of the at-risk sample was further refined to children with 2 years of low problem-solving scores on district-wide tests but reading scores in the average range. Additionally, children whose total composite problem-solving scores (TOMA, KeyMath) were at borderline (i.e., 26th percentile) as being classified at risk for math difficulties (to be discussed) were excluded from the data analysis. It is also important to note that the labels assigned to the three strategy conditions were primarily derived from what was emphasized (e.g., diagrams were utilized in the visual condition on the assumption they were creating an external problem presentation), and therefore it is important to note that elements of both verbal and visual information occur in all the conditions. The verbal treatment conditions drew strategy steps and activities designed to cue attention based on the work of Montague, Warger, and Morgan (2000); Fuchs et al. (2004), and Jitendra et al. (1998), whereas the visual-spatial intervention drew upon the work of van Garderen (2007) and related studies using diagrams from the Singapore curriculum (e.g., Kolloffel et al., 2009; Looi & Lim, 2009; Ng & Lee, 2009).



Table 1

*Classification and Pretest Scores for Children With MD and Average Achievers*

Measure	Reliability	Children with MD			Average achievers			F ratio
		N	M	SD	N	M	SD	
Age		59	8.79	0.75	88	8.78	0.50	0.01
Classification								
TOMA-S	0.87	59	6.17	1.06	88	9.67	2.06	142.68***
KeyM_S	0.90	59	6.66	1.45	88	10.85	2.19	90.63***
Average	0.89	59	6.64	1.09	88	11.16	1.56	371.44***
Fluid intelligence								
Raven_S	0.91	51	97.81	12.83	84	107.61	11.3	21.54**
Reading								
TORC_S	0.98	54	9.52	2.2	84	11.42	1.95	28.16**
WRAT_S	0.81	58	98.55	9.91	88	110.5	11.67	42.30**
Arithmetic								
WIAT_S	0.86	58	94.91	10.65	87	104.44	9.82	30.58**
Working memory								
Concept	0.87	59	2.93	2.06	88	7.5	5.61	35.24**
Sent/Dig	0.86	59	4.79	3.45	88	9.13	5.39	29.47**
Update	0.84	59	3.68	2.49	88	9.02	4.29	74.61**
Composite	0.85	59	-0.49	0.35	88	0.61	0.62	153.61**
Pretest								
Problem solving	0.92	59	5.05	2.26	88	9.69	2.62	123.24**
Calculation	0.93	59	24.02	2.66	88	26.48	3.73	19.10***
Operation span	0.87	58	3.84	3.42	88	5.09	4.67	3.05*

Note. \_S at the end refers to standard or scale score. MD = math difficulties; TOMA = Test of Math Ability; KeyM = KeyMath test; Raven = Raven Colored Matrices Test; TORC = Test of Reading Comprehension; WRAT = Wide Range Achievement Test; WIAT = Wechsler Individual Achievement Test; Concept = conceptual span; Sent/Dig = sentence/digit span; Update = updating measure; Problem solving = word problems solving subtest from the Comprehensive Test of Math Abilities (CMAT); Calculation = arithmetic calculation subtest from the WRAT-3.

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

referenced scores on calculation, reading comprehension, and fluid intelligence were above the 35th percentile.

## Design and Treatment Conditions

**Random assignment.** Children were randomly assigned at the individual level within each classroom either to a control group ( $N = 39$ ) or to one of three treatment conditions: verbal strategies ( $N = 37$ ), verbal + visual strategies (diagramming;  $N = 35$ ), and visual-strategies-only (diagramming;  $N = 36$ ).<sup>2</sup> Although the participating children were randomly assigned to each of the different strategy conditions within classrooms, a number of other controls were built into the implementation of the intervention. For example, to control for the impact of the graduate student tutors who implemented the interventions, all tutors were randomly rotated across days of the week and across treatment conditions, so that no one intervention group received instruction from the same graduate tutor each time (i.e., tutor 1 might have presented Strategy A in the morning time slot on Monday, but then tutor 2 presented the next Strategy A lesson to the same children during that time slot on Wednesday). When comparing demographics of the children randomly assigned to one of the four treatment conditions (verbal-only, verbal + visual, visual-only, control), no significant differences emerged between conditions as a function of MD status,  $\chi^2(3, N = 147) = 1.98, p > .05$ ; gender,  $\chi^2(3, N = 147) = 1.14, p > .10$ ; or chronological age,  $F(3, 147) = 1.47, p > .05$ .

**Common instructional conditions.** All the participants interacted with their peers in their homerooms on tasks and activities related to the district-wide math school curriculum. The school-wide instruction across conditions was the enVisionMATH Learn-

ing Curriculum (Pearson Publishers, 2009). The curriculum included visual representations to show how quantities of a word problem were related and general problem-solving steps. The general problem-solving steps in the teacher manual instructed teachers to have children (a) understand, (b) plan, (c) solve, and (d) look back. An independent evaluation (Resendes & Azin, 2008) indicated in random trials (teachers assigned randomly to treatment or control condition) that gains emerged in Grades 2 to 4, following guidelines outlined by the What Works Clearinghouse (2006) standards, with effect sizes relative to control condition in the 0.20 range. A number of the curriculum's elements were also utilized in this study's treatments (e.g., find the key word). However, in contrast to the school district's required instruction, this study's treatment conditions directly focused on specific components of problem solving over consecutive sessions presented in a predetermined order. In addition, the lesson plans for the experimental condition focused directly on the propositional structure of word problems.

**Experimental conditions.** Each experimental treatment condition included 20 scripted lessons administered over 8 weeks. Each lesson was 30 minutes in duration and was administered three times a week in small groups of four to five children. Lesson administration was done by trained tutors (doctoral-level graduate students and/or master's-level research assistants). Children were

<sup>2</sup> The uneven sample size reflects some small attrition in the sample as well as the removal of children not meeting the operational criteria (e.g., low reading scores) from the data analysis for defining the sample as at risk for MD.

presented with individual booklets at the beginning of the lesson, and all responses were recorded in the booklet. Each lesson within the booklet consisted of four phases: (a) warm-up, (b) strategy instruction, (c) guided practice, and (d) independent practice.

The *warm-up phase* included two parts: calculation of problems that required participants to provide the missing numbers ( $9 + 2 = x$ ,  $x + 1 = 6$ ;  $x - 5 = 1$ ), and a set of puzzles based on problems using geometric shapes. This activity took approximately three to five minutes to complete.

The *instruction phase* lasted approximately five minutes. At the beginning of each lesson, the strategies and/or rule cards were either read to the children (e.g., “to find the whole, you need to add the parts”) or reviewed. Depending on the treatment condition, children were taught the instructional intervention (verbal strategy, diagramming, or verbal strategy + diagramming). The steps for the verbal-strategy-only approach included (a) find the question and underline it, (b) circle the numbers, (c) put a square around the key word, (d) cross out information not needed, (e) decide on what needs to be done (add/subtract/or both), and (f) solve it. For the visual-strategy-only condition (diagramming) children were taught how to use two types of diagrams. The first one represented how parts made up a whole. The second type of diagram represented how quantities are compared. The diagram consisted of two empty boxes, one bigger and the other smaller, in which children were to fill in the correct numbers representing the quantities. An equation with a question mark was presented. The question mark acted as a placeholder for the missing number provided in the box. Finally, for the combined verbal + visual (diagramming) strategy condition, an additional step (diagramming) was added to the six verbal strategy steps described above. This step included directing children to fill in the diagram with given numbers and identify the missing numbers (question) in the corresponding slots in the boxes.

The third phase, *guided practice*, lasted 10 minutes and involved children working on three practice word problems. Tutor feedback was provided on the application of steps and strategies to each of these three problems. In this phase, children also reviewed example problems from the instructional phase. The tutor assisted children with finding the correct operation, identifying the key words, and providing corrective feedback on the solution.

The fourth phase, *independent practice*, lasted 10 minutes and required children to independently (without feedback) answer another set of three word problems. If the child finished the independent practice tasks before the 10 minutes were over, he or she was presented with a puzzle to complete. The child's responses were recorded for each session to assess the application of the intervention and problem-solving accuracy. For the visual-strategy-only condition, points were recorded for choosing the correct diagram, filling in the numbers correctly for the diagram, identifying the correct operations, and solving the problem correctly. For the verbal + visual strategy condition, points were recorded for choosing the correct diagram, inserting correct numbers, applying strategies, identifying the correct operations, and solving the problem correctly. For the verbal-strategy-only condition, points were recorded for identifying the correct numbers, applying strategies (e.g., underlining), identifying the correct operations, and solving the problem accurately.

**Sentence demands.** Word problems for each *independent practice* session included three parts: question sentences, number

sentences, and irrelevant sentences. For each problem in the independent practice session, at least two number sentences were relevant to problem solution and one sentence served as the question sentence. The number of sentences, however, gradually increased across the training sessions. The numbers of sentences were as follows: Lessons 1 through 7 focused on identifying critical information for word problems four sentences long with one irrelevant sentence, Lessons 8 and 9 focused on five-sentence-long word problems with two irrelevant sentences, Lessons 10 through 15 focused on six-sentence-long word problems with three irrelevant sentences, Lessons 16 and 17 focused on seven-sentence-long word problems with four irrelevant sentences, and Lessons 18 through 20 focused on eight-sentence-long word problems with five irrelevant sentences.

**Treatment fidelity.** Independent evaluations were administered in order to determine treatment fidelity. During all lesson sessions, tutors were randomly evaluated by two independent observers (a postdoctoral student, a nontutoring graduate student, and/or the project director). The observers independently filled out evaluation forms covering all segments of the lesson intervention. Points were recorded on the accuracy with which the tutor implemented the instructional sequence based on a rubric. Observations of each tutor occurred for six sessions randomly distributed across instructional sessions. Interrater agreement was calculated on all observation categories. The mean percentage of interrater agreement across all sequences and conditions for each step of strategy implementation (10 observable items were coded) was 98% ( $SD = .41$ ). Mean percent fidelity ratings by strategy conditions were 100.00 ( $SD = 0$ ), 97.05 ( $SD = 4.69$ ), and 97.36 ( $SD = 4.52$ ) for verbal-only, verbal + visual, and visual-only, respectively.

## Tasks and Materials

The battery of group and individually administered tasks is described below. Experimental tasks are described in more detail than published and standardized tasks. Tasks were divided into classification, pretest-only, and pretest/posttest measures. The sample reliabilities (Cronbach alpha) for each measure are shown in Table 1.

## Classification Measures

**Fluid intelligence.** The Raven Colored Progressive Matrices (Raven, 1976) was administered to determine if all children were within the normal range on a measure of fluid intelligence. Children were presented patterns displayed on each page, with each pattern revealing a missing piece. For each pattern, six possible replacement pattern pieces were displayed. Children were required to circle the replacement piece that best completed the pattern. The dependent measure (raw score range 0 to 36) was the number of problems solved correctly, which yielded a standardized score ( $M = 100$ ,  $SD = 15$ ).

**Word problems.** Two measures were administered to assess word problem solving ability. The word problem subtests from the Test of Math Ability (TOMA-2; Brown et al., 1994) and KeyMath (KEYM; Connolly, 1998) were administered. Subtests from these measures yielded a scale score ( $M = 10$ ,  $SD = 3$ ).



## Reading Skills

Several studies have found that working memory is unrelated to word problem solving accuracy when reading proficiency scores are entered in a regression analysis (Ng & Lee, 2009; Swanson et al., 1993). Thus, it was necessary to administer reading measures at pretest because of the potential to moderate treatment outcomes.

**Word recognition.** Word recognition was assessed by the reading subtest of the WRAT-3 (Wilkinson, 1993). The task provided a list of words of increasing difficulty. The child's task was to read the words until 10 errors occurred. The dependent measure was the number of words read correctly.

**Reading comprehension.** Reading comprehension was assessed by the Passage Comprehension subtest from the Test of Reading Comprehension (TORC-III; Brown, Hammill, & Weiderholt, 1995). The purpose of this task was to assess the child's comprehension of topic or subject meaning during reading activities. Comprehension questions were drawn from the reading of short paragraphs. The dependent measure was the number of questions answered correctly.

**Arithmetic calculation.** Because the focus in this study was on problem solving and not calculation per se, I tested whether the children with and without MD were in the normal range on calculation skills. The arithmetic subtest from the Wechsler Individual Achievement Test (WIAT; Psychological Corporation, 1992) was individually administered. This task required the written computation of problems that increased in difficulty. Problems began with simple calculations ( $2 + 2 =$ ) and worked up to more elaborate algebraic calculations. The dependent measure was the number of problems correct, which yielded a standard score ( $M = 100$ ,  $SD = 15$ ).

## Working-Memory Capacity

Several studies have shown individual differences in working memory span (referred to here as working memory capacity; WMC) play a major role in problem-solving performance (e.g., Swanson et al., 2008). Thus, I measured WMC to determine its effect on solution accuracy as a function of treatment conditions. The WMC tasks required children to hold increasingly complex information in memory while responding to a question about the task. The questions served as distractors to item recall because they reflected the recognition of targeted and closely related nontargeted items. A question was asked for each set of items, and the tasks were discontinued if the question was answered incorrectly or if all items within a set could not be remembered. For this study, two WM tasks were administered (conceptual span and sentence/digit task) that followed this format. A separate WM task, referred to as updating, was also administered. The WMC score was the composite  $z$  score that was formed by averaging across the three measures discussed below.

**Conceptual span task.** The purpose of this task was to assess the participant's ability to organize sequences of words into abstract categories (Swanson, 1992, 1995). The participant was presented a set of words (one every 2 seconds), asked a discrimination question, and then asked to recall the words that "go together." For example, a set might include the following words: *shirt, saw, pants, hammer, shoes, nails*. Children were directed to retrieve the words that "go together" (i.e., *shirt, pants, and shoes; saw, hammer, and nails*). The discrimination question was "Which word,

'saw' or 'level,' was said in the list of words?" Thus, the task required participants to transform information encoded serially into categories during the retrieval phase. The range of set difficulty was two categories of two words to five categories of four words. The dependent measure was the highest set recalled correctly (range of 0 to 8) in which the process question was answered correctly.

**Sentence/digit span.** This task assesses the child's ability to remember numerical information embedded in a short sentence (Swanson, 1992, 1995). Before stimulus presentation, the child was shown a card depicting four strategies for encoding numerical information to be recalled. The pictures portrayed the strategies of rehearsal, chunking, association, and elaboration. The experimenter described each strategy to the child before the administration of targeted items. After all strategies were explained, the child was presented numbers in a sentence context. For example, Item 3 states, "Now suppose somebody wanted to have you take them to the supermarket at 8 6 5 1 Elm Street." The numbers were presented at 2-s intervals, followed by a process question (i.e., "What was the name of the street?"). Then, the child was asked to select a strategy from an array of four strategies that represented the best approximation of how he or she planned to practice the information for recall. Finally, the examiner prompted the child to recall the numbers from the sentence in order. No further information about the strategies was provided. Children were allowed 30 seconds to remember the information. Recall difficulty for this task ranged from 3 digits to 14 digits. The dependent measure was the highest set correctly recalled (range = 0–9) in which the process question was answered correctly.

**Updating.** Because WM tasks were assumed to tap a measure of controlled attention referred to as updating (e.g., Miyake, Friedman, Emerson, Witzki, & Howerter, 2000), an experimental updating task, adapted from Morris and Jones (1990), was also administered. A series of one-digit numbers were presented that varied in set lengths of nine, seven, five, and three. No digit appeared twice in the same set. The examiner told the child that the length of each list of numbers might be three, five, seven, or nine digits. Children were then told that they should recall only the last three numbers presented. The digits were presented at approximately 1-second intervals. After the last digit was presented the participant was asked to name the last three digits in order. In contrast to the aforementioned WM measures, which involved a dual-task situation where children answered questions about the task while retaining information (words or spatial location of dots), the current task involved the active manipulation of information such that the order of new information was added to or replaced the order of old information. That is, to recall the last three digits in an unknown ( $N = 3, 5, 7, 9$ ) series of digits, the children must keep the order of old information available (previously presented digits) along with the order of newly presented digits. The dependent measure was the total number of sets correctly repeated (range 0 to 16).

## Pretest and Posttest Measures

**Word problem solving accuracy (CMAT).** Because children were classified as at risk for MD on the TOMA and KeyMath, a separate norm-referenced measure of word problem solving accuracy was individually administered at pretest and posttest: the

Story Problem subtest from the Comprehensive Mathematical Abilities Test (CMAT; Hresko, Schlieve, Herron, Swain, & Sherbenou, 2003). The technical manual for this subtest reported adequate reliabilities ( $>.86$ ) and moderate correlations ( $>.50$ ) with other math standardized tests (e.g., the Stanford Diagnostic Mathematics Test). The test included story problems that increased in solution difficulty. Two forms of the measures varied only in names and numbers. The two forms were counterbalanced across presentation order.

## Transfer

I was interested in how well treatment effects would transfer to other tasks besides the problem solving measure (CMAT). I selected two tasks that I assumed tapped into near transfer and far transfer. The near transfer task (defined as tasks that matched the focus of intervention) focused on calculation. I assumed that because the children in each training session were receiving practice in calculation solution accuracy and that this skill was closely aligned with the intervention, some increases in computation accuracy were to be expected. For the far transfer measure (task not directly related to the focus of treatment), I assessed improvements in working memory on the operation span measure. The measures involved holding in working memory both verbal and computation information of increasing difficulty. Therefore, I assumed that because the cognitive strategy instruction in this study integrated verbal and calculation information, some transfer on the aforementioned measures might occur.

**Calculation.** The arithmetic subtest from the WRAT-3 (Wilkinson, 1993) was individually administered. Two forms of the tests were counterbalanced across children at pretest and posttest. The subtests required written computation to problems that increased in difficulty.

**Operation span.** A version of the Turley-Ames and Whitfield (2003) operation span task, modified for children (Swanson, Kehler, & Jerman, 2010), was individually administered at pretest and posttest. Two identical forms were created and counterbalanced for presentation order. The operation span test assessed WM span by having children solve simple math problems (e.g.,  $2 + 3 =$ ,  $4 - 1 =$ ) while also remembering unrelated to-be-remembered words (e.g., *car*, *pencil*) that followed each math problem. Operation-word sequences increased in set size. Children completed two practice trials with a set size of two. Children were then presented with operation-word sequences in sets of two, three, four, and five, with two trials for each set size for a total of 10 sets. Two versions of test stimuli (form A and form B) were counterbalanced for presentation order. Children received points toward their span score for correctly solving the math problems, for the number of correctly recalled words, and for correct order of word recall.

## Statistical Analyses

Children were drawn from 22 third-grade classrooms. Because the data reflected treatments for children within classrooms, a mixed analysis of covariance (ANCOVA) model was necessary to analyze treatment effects. The fixed and random effect parameter estimates were obtained using PROC MIXED in SAS 9.3. The primary model used in this study was a 2 (MD vs. average

achievers)  $\times$  4 (treatment) mixed ANCOVA. The covariates for the mixed ANCOVA were the continuous variables of pretest and working memory capacity.

All four treatments were administered within each of the 22 classrooms. Within each classroom, children were randomly assigned to treatments. Tutors ( $N = 17$ ) were crossed across classrooms and treatments. That is, except in the control condition, all tutors took turns administering each treatment condition within each classroom. This rotation of tutors was done to ensure that posttest outcomes were related to the treatment procedures rather than the tutors assigned to administer the treatments. The formula for the cross-classification intercept-only model was as follows (see Hox, 2010, Chapter 9, for a review):

$$Y_{i(jk)} = \beta_{o(jk)} + e_{i(jk)} \quad e_{i(jk)} \sim N(0, \sigma^2) \quad (1)$$

where  $Y_{i(jk)}$  is problem-solving accuracy at the posttest for pupil  $i$  within the cross-classification of tutor ( $j$ ), classroom ( $k$ ), and treatments was modeled by the intercept (the overall mean)  $\beta_{o(jk)}$  for the specific combination of tutor and classroom and a residual error term  $e_{i(jk)}$ . The subscripts ( $jk$ ) are written within parentheses to indicate they are conceptually at the same level: the  $jk$  tutor/classroom combination in the cross-classification of tutor and classroom. The subscripts ( $jk$ ) indicated that the intercept  $\beta_{o(jk)}$  varied independently across both tutor and classroom. The error term,  $e_{i(jk)}$ , reflects the deviation of the child's  $i(jk)$  score from the cell mean. These deviations were assumed to be normally distributed with mean 0 and a within-cell variance  $\sigma^2$ .

Thus, I modeled the intercepts with the second-level equation:

$$\beta_{o(jk)} = \gamma_{00} + v_{oj} + v_{ok} \quad (2)$$

In Equation 2,  $v_{oj}$  is the residual term for tutors, and  $v_{ok}$  is the residual term for the classroom. After substitution, this produces the intercept-only model as

$$Y_{i(jk)} = \gamma_{00} + v_{oj} + v_{ok} + e_{i(jk)} \quad (3)$$

I added to this model (Equation 3) the categorical variable of treatment and math ability as well as the continuous variables of pretest and WMC scores as covariates as well as the interactions.

$$\begin{aligned} Y_{i(jk)} = & \gamma_{00} + \gamma_{10}(\text{treatment}) + \gamma_{20}(\text{math ability}) + \gamma_{30}(\text{pretest}) \\ & + \gamma_{40}(\text{WMC}) + \gamma_{40}(\text{WMC} * \text{treatment}) \\ & + \gamma \dots \dots (\text{two-way and three-way interactions}) \\ & + v_{oj} + v_{ok} + e_{i(jk)}. \end{aligned} \quad (4)$$

The intraclass correlations when predicting posttest scores with only random effects (tutor and classroom  $\times$  tutor) were as follows:  $\rho_s = .15$  ( $\tau_0^2 = 0$ ,  $\tau_1^2 = .17$ ,  $\sigma^2 = .92$ ) for problem-solving accuracy,  $\rho_s = .35$  ( $\tau_0^2 = .001$ ,  $\tau_1^2 = .55$ ,  $\sigma^2 = .98$ ) for calculation accuracy, and  $\rho_s = 0$  ( $\tau_0^2 = 0$ ,  $\tau_1^2 = \text{negligible}$ ,  $\sigma^2 = .93$ ) for operation span. The intraclass correlations for predicting posttest scores were reduced to 0 for problem-solving accuracy, .15 for calculation accuracy, and 0 for operation span when treatment conditions, ability group, WMC, pretest scores, and interactions were entered into the full model.

Because the cells are unbalanced, a Kenward–Roger correction was used to obtain degrees of freedom. A full maximum-likelihood



(ML) estimation was used to compute the parameters at posttest because of some attrition in sample size (Widaman, 2006).

## Results

Table 1 provides the means and standard deviations for the classification and criterion (pretest) measures. The  $F$  ratios comparing the ability groups prior to treatment assignment and the sample reliability of the measures are also reported in Table 1. Sample sizes, pretest scores, and posttest scores for children with and without MD as a function of treatment conditions are reported in Table 2.

### Pretest

**Criterion measures.** Because children were randomly assigned to each condition within classrooms, it was necessary to determine if potential biases in treatment assignment emerged at pretest. The criterion measures used to assess treatment effects were word problems from the CMAT, arithmetic problems from the WRAT-3, and recall scores from the operation span measure. Equivalent forms were developed for each measure, and the presentation orders were counterbalanced across treatment conditions.

A 2 (MD vs. average achievers)  $\times$  4 (treatment condition) mixed analysis of variance (ANOVA) was computed on pretest scores. The random effects included variance related to the aforementioned intercepts for tutors and classroom assignment. As expected, the main effect was significantly in favor of average achieving children when compared to children with MD on measures of pretest problem-solving accuracy,  $F(1, 147) = 129.44$ ,  $p < .001$ ; pretest calculation accuracy,  $F(1, 147) = 18.89$ ,  $p < .001$ ; and pretest operation span accuracy performance,  $F(1, 143) = 4.80$ ,  $p = .03$ . The main effects for treatment conditions were not significant, however, for pretest problem solving,  $F(3, 147) = 0.59$ ,  $p = .62$ ; pretest calculation,  $F(3, 147) = 0.75$ ,  $p = .52$ ; or pretest operation span performance,  $F(3, 143) = 1.46$ ,  $p = .33$ . In addition, no significant effects occurred for the ability group  $\times$  treatment interactions on measures of pretest problem solving,  $F(3, 147) = 2.07$ ,  $p = .11$ ; pretest calculation,  $F(3, 147) = 0.75$ ,  $p = .52$ ; or pretest operation span performance,  $F(3, 141) = 2.06$ ,  $p = .11$ . A mixed 2 (ability group)  $\times$  4 (treatment) ANOVA was also computed on the WMC composite scores. A significant effect was found for ability group,  $F(1, 147) = 161.08$ ,  $p < .001$ , but not for the main effect of treatment,  $F(3, 147) = 0.92$ ,  $p = .43$ , or for the ability group  $\times$  treatment interaction,  $F(3, 147) = 0.31$ ,  $p = .82$ .

**Classification.** The two ability groups were compared across treatment conditions on the classification measures. A 2 (ability group = MD vs. average achievers)  $\times$  4 (treatment) multivariate analysis of variance (MANOVA) was computed on classification measures of problem solving (TOMA, KeyMath), reading (TORC), fluid intelligence (Raven Colored Matrices Test), and arithmetic calculation (WIAT). The MANOVA was significant for ability group, Wilks's  $\Lambda = .48$ ,  $F(4, 114) = 19.00$ ,  $p < .001$ , but not for treatment, Wilks's  $\Lambda = .74$ ,  $F(12, 301) = 1.27$ ,  $p = .23$ , or for the ability group  $\times$  treatment interaction, Wilks's  $\Lambda = .79$ ,  $F(12, 301) = 1.17$ ,  $p = .30$ . As expected, all univariate tests of significance were statistically significant in favor of the average achievers when compared to children with MD (see Table 1; all  $ps <$

.05). It is important to note, however, that although fluid intelligence, reading, and calculation scores were in the normal range for children with MD, average achieving children yielded higher scores than children with MD on these measures (see Table 1).

For the next series of analyses, posttest criterion measures were converted to  $z$  scores based on the total sample means and standard deviations at pretest. This conversion allowed for comparisons across various dependent measures, as well as the identification of outliers (absolute  $z$  score  $> 3.5$ ). No outliers were identified.

### Posttest CMAT Solution Accuracy

A 2 (ability group)  $\times$  4 (treatment) mixed ANCOVA was computed on posttest  $z$  scores. The covariates for the analyses were the continuous variables of pretest CMAT solution accuracy and WMC. The mixed ANCOVA yielded significant effects for the ability group  $\times$  treatment interaction,  $F(1, 147) = 3.21$ ,  $p = .02$ ; the WMC  $\times$  treatment interaction,  $F(3, 147) = 8.10$ ,  $p < .001$ ; the ability group  $\times$  treatment  $\times$  WMC interaction,  $F(3, 147) = 3.51$ ,  $p = .02$ ; and the pretest CMAT score,  $F(1, 147) = 205.71$ ,  $p < .001$ . No other significant effects occurred ( $ps > .05$ ). For example, no significant main effects occurred for ability group,  $F(1, 147) = 1.42$ ,  $p = .23$ , or treatment condition,  $F(3, 147) = 0.74$ ,  $p = .53$ . The adjusted posttest  $z$  score means for each treatment condition as a function of ability group are shown in Table 3.

Because of the significant ability group  $\times$  treatment  $\times$  WMC interaction, a series of follow-up tests was conducted. The first follow-up determined if the slopes varied significantly between treatments as a function of ability group. The slopes for each treatment as a function of ability group are shown in Table 3. As shown, slopes were significantly higher for average achieving children when compared to children with MD within the verbal + visual condition,  $t(147) = -1.96$ ,  $p = .05$ , and the control condition,  $t(147) = -2.57$ ,  $p = .01$ . However, slopes were significantly higher for children with MD when compared to average achieving children within the visual-only condition,  $t(147) = 2.40$ ,  $p = .02$ . No significant ability group differences in slopes occurred within the verbal-only condition,  $t(147) = 1.01$ ,  $p = .31$ . The slopes for the control condition were next compared to the other treatment conditions within each ability group. When compared to those for children with MD in the control condition, the slopes were significantly larger for children with MD within the verbal-only condition,  $t(147) = 2.48$ ,  $p = .01$ , and visual-only condition,  $t(147) = 3.59$ ,  $p < .001$ , but not when compared to children with MD within the verbal + visual condition,  $t(147) = .14$ ,  $p = .88$ . When compared to those for average achieving children in the control condition, the slopes were significantly larger for average achieving children within the verbal-only condition,  $t(147) = 2.64$ ,  $p = .009$ ; verbal + visual condition,  $t(147) = 2.24$ ,  $p = .03$ ; and the visual-only condition,  $t(147) = -3.66$ ,  $p = .008$ .

Figures 1a and 1b show the linear regression line for each treatment condition as a function of WMC on the adjusted posttest solution accuracy scores for children with and without MD, respectively. As shown, posttest scores as a function of treatment (i.e., verbal-only and visual-only treatment conditions) were clearly divergent from the control condition, as WMC  $z$  scores approached approximately  $-1.0$  for children with MD (see Figure 1a), whereas clear treatment divergence from the control condition

Table 2  
Means and Standard Deviations as a Function of Treatment and Ability Groups

Variable	Verbal-only			Verbal + visual			Visual-only			Control		
	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>
Children with MD												
Age	12	8.79	0.79	17	8.86	0.70	14	8.55	0.52	16	8.92	0.93
Classification												
TOMA-S	12	5.92	1.16	16	6.06	1.29	14	6.21	0.89	16	6.44	0.89
KeyM-S	12	7.00	0.89	17	6.44	0.73	14	6.25	2.05	16	7.60	0.89
Math	12	6.42	1.38	17	6.47	1.12	14	6.79	0.8	16	6.88	1.09
Fluid intelligence												
Raven-S	10	95.97	19.75	14	100.58	12.13	13	97.27	11.41	14	96.85	9.15
Reading												
TORC-S	10	10.70	1.77	17	9.00	1.80	12	9.50	2.81	15	9.33	2.26
WRAT-S	12	100.00	7.69	17	98.18	12.53	13	98.77	9.96	16	97.69	8.96
Arithmetic												
WIAT-S	12	98.00	11.88	17	92.65	12.45	13	95.08	8.39	16	94.88	9.63
Working memory												
Concept_R	12	2.00	1.28	17	3.71	2.64	13	3.00	1.15	16	2.75	2.24
Sent/Dig_R	12	5.08	3.42	17	5.47	3.26	13	4.54	3.5	16	4.06	3.79
Update_R	12	4.58	3.42	17	3.35	1.73	14	3.00	1.92	16	3.94	2.77
WMC	12	-0.47	0.32	17	-0.40	0.30	14	-0.60	0.45	16	-0.52	0.31
Pretest												
Problem solving	12	5.00	1.81	17	4.65	1.93	14	4.50	2.62	16	6.00	2.45
Calculation_R	12	24.50	3.21	17	23.47	3.14	14	23.14	1.7	16	25.00	2.13
Operation span_R	12	1.92	1.44	16	4.38	3.76	14	4.36	3.95	16	4.31	3.38
Posttest												
Problem solving	12	6.33	2.10	17	6.47	2.83	13	5.62	3.5	16	7.25	2.59
Calculation_R	12	26.00	2.70	17	25.65	2.50	13	24.46	3.36	16	26.44	3.29
Operation span_R	12	3.83	3.01	17	5	3.08	14	6.36	4.01	16	5.75	2.86
Average achievers												
Age	25	8.77	.60	18	8.88	.50	22	8.66	.36	22	8.83	.48
Classification												
TOMA-S	25	9.88	2.57	18	9.56	1.69	22	9.32	1.96	22	9.82	1.87
KeyM-S	25	10.77	2.17	18	9.67	1.86	22	12.07	1.98	22	10.15	2.23
Math_S	25	11.40	1.41	18	10.28	1.23	22	11.91	1.69	22	10.86	1.52
Fluid intelligence												
Raven-S	24	110.91	9.32	16	110.05	8.95	22	107.83	10.56	21	101.5	13.93
Reading												
TORC-S	24	11.92	2.00	17	11.18	1.70	21	11.57	1.63	21	11.19	1.91
WRAT-S	25	108.84	11.06	18	110.94	10.26	22	112.32	11.81	22	110.23	13.88
Arithmetic												
WIAT-S	25	104.60	11.53	17	102.53	8.25	22	107.05	6.83	22	103.05	11.49
Working memory												
Concept_R	25	5.60	3.99	18	6.89	3.14	22	9.86	7.63	22	7.86	5.95
Sent/Dig_R	25	10.08	5.69	18	10.67	5.65	22	7.27	4.7	22	8.68	5.32
Update_R	25	9.48	4.81	18	10.44	3.81	22	8.23	4.0	22	7.91	4.12
WMC <sup>a</sup>	25	0.54	0.62	18	0.77	0.57	22	0.63	0.77	22	0.52	0.51
Pretest												
Problem solving_R	25	10.48	2.35	18	9.50	2.81	22	9.73	2.43	22	8.91	2.91
Calculation_R	25	27.08	4.05	18	26.11	4.74	22	26.59	3.11	22	25.95	3.20
Operation span_R	25	5.64	4.69	18	6.72	5.85	22	3.95	3.9	22	4.50	4.10
Posttest												
Problem solving_R	25	10.4	2.90	18	10.83	2.75	20	10.5	1.93	19	9.68	2.38
Calculation_R	25	29	3.21	18	28.44	4.2	20	27.75	2.92	20	29.9	2.88
Operation span_R	25	7.32	4.23	18	8.00	5.64	22	7.00	3.70	22	6.09	4.03

Note. \_R at the end refers to raw score; \_S at the end refers to standard or scale score. TOMA = Test of Math Ability; KeyM = KeyMath test; Math\_S = mean scale-score (TOMA, KeyM); Raven = Raven Colored Matrices Test; TORC = Test of Reading Comprehension; WRAT = Wide Range Achievement Test; WIAT = Wechsler Individual Achievement Test; Concept = conceptual span; Sent/Dig = sentence/digit span; Update = updating measure; WMC = working memory capacity; Problem solving = word problems solving subtest from the Comprehensive Test of Math Abilities (CMAT); Calculation = arithmetic calculation subtest from the WRAT-3.

<sup>a</sup> Denotes composite mean *z* score of working memory span measures (conceptual span, digit/sentence span, and updating).



Table 3  
*Estimated Adjusted Posttest Z Scores and Slopes for Problem Solving, Calculation, and Operation Span as a Function of Treatment and Ability Group*

Group	Problem solving		Calculation		Operation span	
	Adjusted <i>M</i>	<i>SE</i>	Adjusted <i>M</i>	<i>SE</i>	Adjusted <i>M</i>	<i>SE</i>
Posttest accuracy						
Verbal-only						
MD	0.99	0.29	1.28	0.42	0.54	0.37
AVE	0.67	0.11	1.4	0.16	0.53	0.13
Verbal + visual						
MD	0.26	0.24	1.46	0.34	0.28	0.31
AVE	0.78	0.16	0.97	0.23	0.49	0.19
Visual-only						
MD	1.04	0.24	1.94	0.35	1.09	0.31
AVE	0.72	0.11	1.26	0.17	0.89	0.14
Control						
MD	0.16	0.27	0.88	0.40	0.05	0.35
AVE	0.91	0.12	2.2	0.17	0.16	0.14
Slopes						
Verbal-only						
MD	0.69	0.42	0.38	0.59	0.22	0.52
AVE	0.25	0.15	0.2	0.21	0.04	0.19
Verbal + visual						
MD	-0.61	0.37	0.41	0.54	0.05	0.51
AVE	0.21	0.19	0.78	0.28	0.01	0.24
Visual-only						
MD	0.96	0.28	1.4	0.39	0.7	0.35
AVE	0.22	0.13	0.16	0.18	-0.25	0.16
Control						
MD	-0.68	0.37	-0.44	0.53	-0.52	0.47
AVE	-0.38	0.19	-0.28	0.27	0.74	0.24

Note. MD = children with math difficulties; AVE = children without math difficulties, or average achieving children.

occurred for average achieving children when the WMC *z* score was at approximately 1.0 (see Figure 1b). Figures 1a and 1b also show a point at which treatment and control conditions intersect. As shown in Figure 1a, this intersection point occurred at approximately -0.5 WMC *z* score for children with MD, whereas as shown in Figure 1b this intersection point occurred at approximately 0.5 *z* score for average achieving children. Thus, as a follow-up to the covariate (WMC) by treatment interaction, posttest scores were estimated when made conditional on setting WMC to high (1.0 *z* score), middle (0 *z* score), and low (-1.0 *z* score) values.<sup>3</sup> Pretest CMAT scores again served as a covariate in the analysis. The estimated adjusted mean posttest scores when made conditional on setting WMC to high, middle, and low values are reported in Table 4.

At the high WMC level (1.0 WMC values), significant treatment effects occurred at posttest for children with MD,  $F(3, 147) = 4.99$ ,  $p = .003$ , but not for average achieving children,  $F(3, 147) = 1.85$ ,  $p = .14$ . For children with MD, a Tukey test indicated a significant advantage at posttest for the visual-only and verbal-only condition compared to the other treatment conditions (visual-only = verbal-only > verbal + visual = control). A significant posttest advantage was found for average achieving children when compared to children with MD within the verbal + visual condition,  $F(1, 147) = 4.72$ ,  $p < .04$ . In contrast, children with MD outperformed average achieving children at posttest within the visual-only condition,  $F(1, 147) = 4.05$ ,  $p < .045$ . No other significant effects ( $ps > .05$ ) occurred in estimated

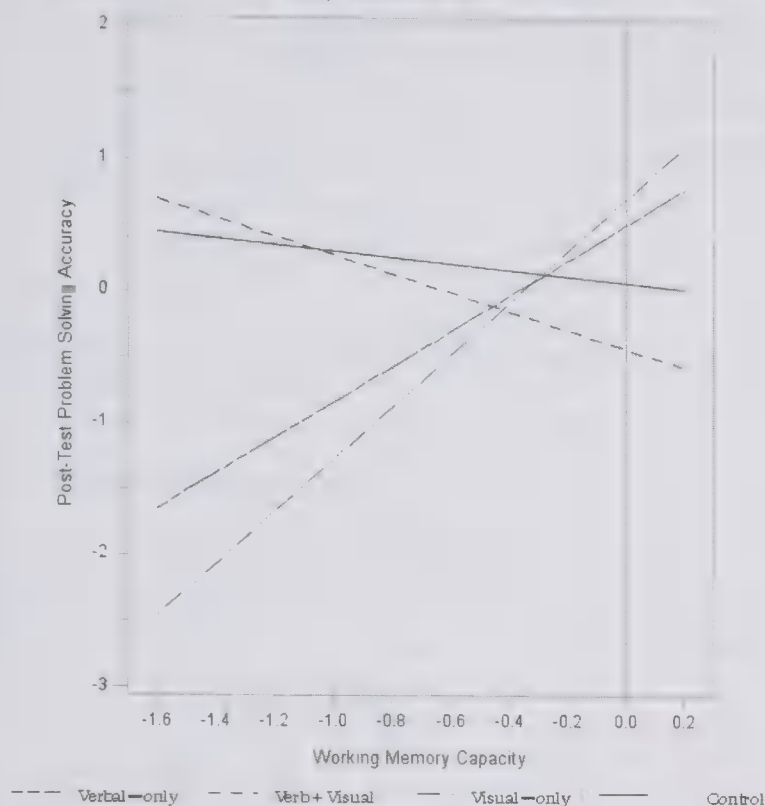
posttest scores when made conditional on setting WMC to a high (1.0 *z* score) level.

At the middle WMC level (0 *z* score), no significant treatment effects occurred for children with MD,  $F(3, 147) = 2.39$ ,  $p = .07$ , or for average achieving children,  $F(3, 147) = 1.23$ ,  $p = .29$ . The only significant ability group difference in posttest performance

<sup>3</sup> I did not compare treatment differences at each point of the WMC covariate. For parsimony, I selected a cutoff point (referred to as a pick a point approach; Rogosa, 1980) for comparisons at the lines where treatment outcomes started to diverge in Figure 1a and Figure 1b. Not unlike when using the Johnson–Neyman procedure (Rogosa, 1980), I did consider potential regions of significance. I initially utilized Bauer and Curran's (2005) procedure (see Lazer & Zerbe, 2011, for SAS syntax) for computing the Johnson–Neyman technique for mixed models. However, because there was sufficient information to conclude the slopes were not equal across all conditions and my hypotheses were tied to variations in outcomes related to high and low WMC, I tested whether the adjusted posttest treatment outcomes depended on whether WMC was set at high, middle, and low values. This procedure, referred to as a treatment by covariate interaction design (e.g., Judd, McClelland, & Smith, 1996; Leon, Portera, Lowell, & Rheinheimer, 1998; Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2010), has the advantage of testing whether the adjusted posttest scores as a function of treatment are conditional on the level at which WMC is set. The SAS syntax for computing the estimated adjusted means (least square means) at posttest conditional on setting the covariate to specific values is provided in Littell et al. (2010, p. 265).

**a Regression Line for WMC and Strategy Training**

Group=Children with MD

**b Regression Line for WMC and Strategy Training**

Group=Average Achiever

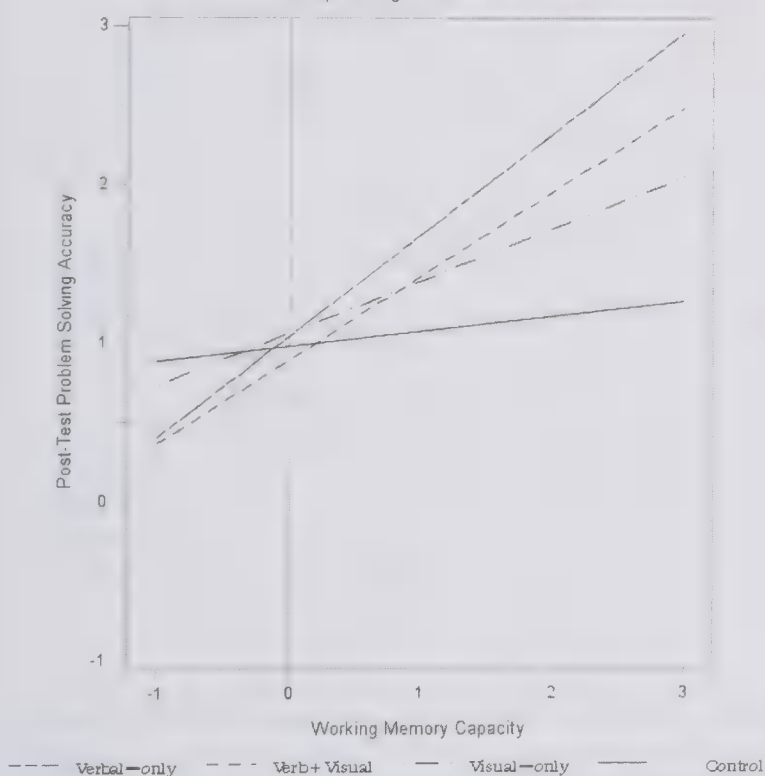


Figure 1. a: Linear regression slope for Treatment  $\times$  Working Memory Capacity (WMC) for children with math difficulties. b: Linear regression slope for Treatment  $\times$  Working Memory Capacity for average achievers. MD = math difficulties.

that occurred was within the control condition. Average achieving children yielded significantly higher adjusted posttest scores than did children with MD,  $F(1, 147) = 7.24, p = .008$ .

At the low WMC level ( $-1.00$   $z$  score), a significant treatment effect occurred for children with MD,  $F(3, 147) = 7.12, p < .001$ , but not for average achieving children,  $F(3, 147) = 2.40, p = .07$ . For children with MD, a Tukey test indicated a posttest advantage ( $ps < .05$ ) for the control and verbal + visual conditions when compared to the other conditions (control = verbal + visual > verbal-only = visual-only). No significant posttest score advantage occurred for children with or without MD within treatment conditions ( $ps > .05$ ).

**Summary.** The results clearly showed that WMC moderated treatment outcomes on measures of posttest solution accuracy. For children with MD with relatively high WMC, an estimated posttest treatment advantage was found for the verbal-only and visual-only conditions when compared to the control condition. In contrast, when WMC was set to a low level, none of the treatment conditions exceeded the control condition at posttest.

For average achievers, no treatment advantages relative to the control condition occurred at posttest. Although the results showed higher slopes for the treatment conditions when compared to the control condition, none of the treatment conditions significantly improved posttest scores when compared to the control condition. These nonsignificant effects held regardless of whether WMC was set to high, middle, or low WMC values.

Taken together, there is weak support for the compensatory hypothesis, which suggests that children with lower WMC are more likely to benefit from strategy conditions when compared to the control condition than are those with higher WMC values.

### Posttest Calculation Accuracy

Because calculation practice was part of the intervention, I expected some improvements in arithmetic skills. The general analytic strategy as used before tested whether there were improvements in calculation accuracy as a function of ability group and treatment conditions. A 2 (ability group)  $\times$  4 (treatment) mixed ANCOVA was computed on posttest calculation (WRAT-3)  $z$  scores. The covariates were the continuous variables of pretest calculation accuracy and WMC. The mixed ANCOVA yielded significant effects for the ability group  $\times$  treatment interaction,  $F(1, 144) = 5.24, p = .002$ ; WMC  $\times$  treatment interaction,  $F(3, 141) = 3.32, p = .02$ ; WMC,  $F(1, 145) = 4.58, p = .03$ ; and the pretest WRAT-3 score,  $F(1, 141) = 164.64, p < .001$ . No other significant effects occurred ( $ps > .05$ ). In particular, no significant main effects occurred for ability group,  $F(1, 147) = .31, p = .58$ ; treatment conditions,  $F(3, 124) = 1.18, p = .32$ ; or the ability group  $\times$  treatment  $\times$  WMC interaction,  $F(3, 129) = 2.20, p = .09$ . The adjusted posttest  $z$  score means for calculation accuracy are shown in in Table 3.

A test of simple effects as a follow-up to the ability group  $\times$  treatment interaction indicated that treatment effects in adjusted posttest scores were significant for average achieving children,  $F(3, 145) = 8.27, p < .0001$ , but not for children with MD,  $F(3, 145) = 1.37, p = .25$ . For average achieving children, a Tukey test yielded a significant ( $ps < .05$ ) adjusted posttest score advantage for the control condition when compared to the treatment conditions (control > visual-only = verbal-only > verbal + visual). Within treatment conditions, a significant advantage in adjusted posttest scores was found for average achieving children when compared to children with MD for the control condition,  $t(147) =$



Table 4  
*Estimated Posttest Z Scores for Problem Solving, Calculation, and Operation Span as a Function of Treatment and Working Memory Capacity Set to High, Middle, and Low Values*

Group	Problem solving		Calculation		Operation span	
	Adjusted <i>M</i>	<i>SE</i>	Adjusted <i>M</i>	<i>SE</i>	Adjusted <i>M</i>	<i>SE</i>
High WMC						
Verbal-only						
MD	1.56	0.62	1.6	0.89	0.72	0.78
AVE	0.88	0.12	1.56	0.17	0.56	0.14
Verbal + visual						
MD	-0.25	0.54	1.81	0.77	0.32	0.72
AVE	0.96	0.12	1.62	0.17	0.51	0.15
Visual-only						
MD	1.85	0.46	3.11	0.65	1.66	0.58
AVE	0.91	0.11	1.39	0.17	0.68	0.13
Control						
MD	-0.41	0.57	0.51	0.83	-0.38	0.72
AVE	0.59	0.13	1.97	0.19	0.78	0.16
Middle WMC						
Verbal-only						
MD	0.87	0.23	1.22	0.34	0.5	0.3
AVE	0.63	0.12	1.37	0.18	0.52	0.15
Verbal + visual						
MD	0.36	0.19	1.4	0.27	0.27	0.24
AVE	0.74	0.18	0.84	0.27	0.49	0.22
Visual-only						
MD	0.89	0.2	1.71	0.3	0.97	0.26
AVE	0.69	0.13	1.23	0.18	0.93	0.16
Control						
MD	0.27	0.22	0.95	0.33	0.14	0.28
AVE	0.97	0.14	2.24	0.2	0.03	0.17
Low WMC						
Verbal-only						
MD	0.18	0.26	0.84	0.36	0.28	0.32
AVE	0.39	0.24	1.17	0.34	0.48	0.31
Verbal + visual						
MD	0.97	0.25	0.98	0.37	0.22	0.35
AVE	0.53	0.35	0.06	0.52	0.48	0.44
Visual-only						
MD	-0.08	0.18	0.31	0.24	0.27	0.2
AVE	0.46	0.22	1.07	0.32	1.18	0.28
Control						
MD	0.96	0.21	1.39	0.3	0.66	0.26
AVE	1.35	0.3	2.52	0.44	-0.71	0.38

*Note.* WMC = working memory capacity; MD = children with math difficulties; AVE = children without math difficulties, or average achieving children; High WMC = estimated adjusted posttest score by setting to a WMC value of 1.0; Middle WMC = estimated adjusted posttest score by setting to a WMC *z*-score value of 0; Low WMC = estimated adjusted posttest score by setting to a WMC value of -1.0.

9.31,  $p = .003$ . No other significant effects (all  $ps > .05$ ) emerged comparing the adjusted posttest scores.

As shown in Table 3, when slopes were compared between ability groups, a larger slope was found for children with MD when compared to average achieving children within the visual-only condition,  $t(128) = 2.91$ ,  $p < .004$ . No other significant differences ( $ps > .05$ ) in slopes occurred between ability groups within conditions. Because no significant ability group  $\times$  treatment  $\times$  WMC interaction occurred, slopes were collapsed across ability groups for the comparison between treatment conditions. When compared to those for the control condition, slopes were significantly higher for the verbal + visual condition,  $t(147) = 2.22$ ,  $p = .02$ , and visual-only condition,  $t(146) = 3.07$ ,  $p = .003$ . No significant difference in slopes occurred comparing the control condition to the verbal-only condition,  $t(143) = 1.49$ ,  $p = .14$ .

As a follow-up to the significant WMC  $\times$  treatment interaction, I again set WMC to high (1.0), middle (0), and low (-1.0) values. The estimated posttest scores as a function of WMC values are shown in Table 4. These values were again selected so comparisons in posttest outcomes could be made across the three dependent measures. Because no significant group  $\times$  treatment  $\times$  WMC interaction emerged, comparisons between treatments focused on the total sample as a function of WMC values. No significant treatment differences (all  $ps > .05$ ) occurred in the estimated adjusted posttest scores as a function of treatment conditions at high and middle WMC values ( $z$  scores of 1.0 and 0). In contrast, when WMC was set to low values (-1.0  $z$  score), children in the control condition performed better than those in the treatment conditions. A Tukey test ( $p < .05$ ) yielded a significant advantage in adjusted posttest scores for the control when compared to the

treatment conditions (control > verbal-only = verbal + visual = visual-only).

**Summary.** No support was found for the notion that, relative to the control condition, strategy conditions provided additional advantages in posttest calculation performance for children with MD. Likewise, no posttest advantages as a function of strategy conditions were found for average achievers relative to the control condition. Regardless of ability group, when WMC was set to high and middle levels, no treatment advantages occurred relative to the untreated control condition. In fact, when WMC was set to a low level, children in the control condition actually performed better at posttest than did those in the treatment conditions.

### Posttest Operation Span

Because strategy interventions included practice with word problems that gradually increased interference or distraction during the training sessions (the number of irrelevant sentences across sessions was gradually increased), I expected that this activity, coupled with the strategy instruction, played an important role in treatment outcomes on working memory measures. Because WMC is defined as including the inhibition of distracting information (e.g., Engle et al., 1999), I tested whether some transfer effects occurred on the operation span measure. A 2 (ability group)  $\times$  4 (treatment) mixed ANCOVA was computed on operation span posttest  $z$  scores. The covariates were pretest operation span and WMC. Both of these covariates were continuous variables. The mixed ANCOVA yielded significant effects for treatment,  $F(3, 146) = 5.16, p = .002$ ; for the ability group  $\times$  treatment  $\times$  WMC interaction,  $F(3, 146) = 3.77, p = .01$ ; and for the pretest operation span score,  $F(1, 146) = 224.82, p < .0001$ . No other significant effects occurred (all  $ps > .05$ ). The adjusted posttest means as well as slopes as a function of group and treatment are shown in Table 3.

As in the previous analyses, a comparison was made between slopes within treatment conditions as a function of ability group. As shown in Table 3, slopes were significantly higher for children with MD than for average achieving children within the visual-only condition,  $t(147) = 2.46, p = .01$ . No other significant group differences ( $ps > .05$ ) in slopes occurred within the remaining treatment conditions (all  $ps > .05$ ). I next compared the slopes of the control condition with those of the other treatment conditions within each ability group. When compared to those for children with MD in the control condition, the slopes were significantly larger for children with MD in the visual-condition,  $t(146) = 2.07, p = .04$ . No other significant slope differences (all  $ps > .05$ ) occurred between the control and treatment conditions within the MD group. When compared to those for average achieving children in the control condition, the slopes were significantly smaller for average achieving children in the verbal-only condition,  $t(146) = -2.30, p = .02$ ; verbal + visual condition,  $t(146) = -2.14, p = .03$ ; and visual-only condition,  $t(146) = -3.47, p < .0001$ .

To follow up on the WMC by treatment by ability group interaction, I again set WMC to high (1.0), middle (0), and low ( $-1.0$ ) values. At the higher WMC values, no significant treatment effects occurred for children with MD,  $F(3, 146) = 1.75, p = .15$ , or for average achieving children,  $F(3, 146) = 0.63, p = .59$ .

Further, no significant effects occurred between groups within treatment conditions (all  $ps > .05$ ).

At the middle WMC level, a significant treatment effect occurred for average achieving children,  $F(3, 146) = 5.00, p < .003$ , but not for children with MD,  $F(3, 146) = 1.95, p = .12$ . For average achieving children, a Tukey test indicated a significant advantage ( $ps < .05$ ) for the visual-only condition relative to the other conditions (visual-only > verbal-only = verbal + visual > control). No significant effects ( $ps > .05$ ) occurred between groups within treatment conditions (all  $ps > .05$ ).

At the low level of WMC ( $-1.0$   $z$  score), a significant treatment effect occurred for average achieving children,  $F(3, 146) = 5.23, p < .002$ , but not for children with MD,  $F(3, 146) = 0.57, p = .63$ . For average achieving children, a Tukey test yielded a significant adjusted posttest advantage ( $ps < .05$ ) for the visual-only condition relative to the other conditions (visual-only > verbal-only = verbal + visual > control). Within conditions, a significant estimated posttest advantage occurred for average achieving children when compared to children with MD for the visual-only condition,  $t(146) = 6.71, p = .01$ . In contrast, the estimated adjusted posttest scores were higher for children with MD than for average achieving children in the control condition,  $t(146) = 8.68, p = .004$ . No other significant group effects on estimated adjusted posttest scores occurred at low WMC level.

**Summary.** For average achieving children and WMC values in the low and middle range, a significant treatment advantage was found for the visual-only condition when compared to the other conditions. For children with MD, no clear treatment advantages in adjusted posttest operation span scores occurred when compared to those in the control condition.

### Effect Sizes

The above statistical outcomes for the various treatment conditions clearly were related to the power in my analysis. Thus, to partially address this issue, I report effect sizes (ESs) in Table 5. I calculated Hedges's  $g = \gamma / [(n_1 - 1)(SD_1^2) + (n_2 - 1)(SD_2^2) / (n_1 + n_2 - 2)]^{1/2}$ , where  $\gamma$  is the hierarchical linear modeling coefficient for the intervention effect, which represents the mean difference between treatment adjusted for both Level 1 and Level 2 covariates;  $n_1$  and  $n_2$  are the sample sizes; and  $SD_1$  and  $SD_2$  are the unadjusted posttest standard deviations (What Works Clearinghouse, 2006; see Formula 10), respectively. The Level 2 coefficients were adjusted for the Level 1 covariates. For the interpretation of the magnitude of the effect sizes (ESs), Cohen's (1988) distinction was used; an ES of 0.20 is considered small, and ESs of 0.50 and 0.80 are considered moderate and large, respectively.

Table 5 shows the magnitude of ESs at posttest for children with MD and average achievers. Reported are the effect sizes for the adjusted posttest scores estimated at high, middle, and low WMC values. Also reported are the adjusted posttest effect sizes when WMC was left to covary. That is, adjusted posttest outcomes are compared without setting WMC to a specific value. Effect sizes comparing the treatment to the control condition that yielded effect sizes at or above .80 are in bold.

For those children with MD and when WMC was set to high values, high effect sizes in favor of verbal-only and visual-visual-only conditions occurred when compared to the control condition on posttest outcome measures of problem solving. The results also



Table 5  
*Effect Sizes for Adjusted Posttest Scores for Problem Solving, Calculation, and Operation Span as a Function of Treatment and Working Memory Capacity at High, Middle, and Low Values*

Group	1 vs. 2	1 vs. 3	1 vs. 4	2 vs. 3	2 vs. 4	3 vs. 4
Children with MD						
Problem solving						
High WMC	2.23	-0.31	<b>2.61</b>	-2.10	0.20	<b>2.35</b>
Middle WMC	0.63	-0.01	0.79	-0.52	0.11	0.64
Low WMC	-0.98	0.28	-1.02	1.06	0.02	<b>-1.07</b>
Sample <sup>a</sup>	0.89	-0.06	<b>1.09</b>	-0.78	0.12	<b>0.92</b>
Calculation						
High WMC	-0.23	-1.47	<b>1.05</b>	-1.31	<b>1.3</b>	<b>2.36</b>
Middle WMC	-0.19	-0.48	0.26	-0.32	0.45	0.69
Low WMC	-0.16	0.52	-0.53	0.68	-0.40	<b>-0.98</b>
Sample <sup>a</sup>	-0.2	-0.64	0.39	-0.48	0.59	<b>0.97</b>
Operation span						
High WMC	0.54	-1.12	<b>1.58</b>	-1.6	<b>0.99</b>	<b>2.50</b>
Middle WMC	0.32	-0.55	0.52	-0.83	0.18	<b>1.01</b>
Low WMC	0.09	0.02	-0.54	-0.06	-0.63	-0.48
Sample <sup>a</sup>	0.36	-0.65	0.70	-0.96	0.32	<b>1.27</b>
Average achievers						
Problem solving						
High WMC	-0.1	-0.04	0.42	0.08	0.57	0.46
Middle WMC	-0.14	-0.07	-0.48	0.09	-0.34	-0.41
Low WMC	-0.18	-0.11	<b>-1.38</b>	0.1	-1.25	<b>-1.28</b>
Sample <sup>a</sup>	-0.13	-0.07	-0.33	0.09	-0.19	-0.26
Calculation						
High WMC	-0.05	0.16	-0.36	0.22	-0.33	-0.52
Middle WMC	0.4	0.12	-0.78	-0.37	-1.32	<b>-0.91</b>
Low WMC	0.84	.09	<b>-1.20</b>	-0.97	-2.32	<b>-1.29</b>
Sample <sup>a</sup>	0.32	0.13	-0.71	-0.28	-1.16	<b>-0.84</b>
Operation span						
High WMC	0.05	-0.13	-0.22	-0.2	-0.29	-0.10
Middle WMC	0.02	-0.43	0.48	-0.49	0.49	<b>0.90</b>
Low WMC	.001	-0.74	<b>1.19</b>	-0.78	1.26	<b>1.89</b>
Sample <sup>a</sup>	0.03	-0.38	0.36	-0.44	0.35	0.73

Effect sizes at or greater than .80 when compared to the control condition are shown in boldface type. Conditions are denoted as follows: 1 = verbal-only, 2 = verbal + visual, 3 = visual-only, 4 = control. Positive effect sizes in favor of the first number (e.g., 1 vs. 2) indicated an advantage for the verbal condition when compared to the verbal + visual (effect size = 2.23). Settings for working memory capacity (WMC) were 1.0 for high WMC, 0 for middle WMC, and -1.0 for low WMC.

<sup>a</sup> Estimates were not made conditional on setting WMC to a specific value.

showed high effect sizes for all three treatment conditions relative to the control condition on estimated posttest measures of calculation accuracy and operation span. The treatment condition that yielded consistently high ESs across all dependent measures when compared to the control condition was the visual-only condition (ESs ranged from 2.35 to 2.50). When WMC was set to low values, the effect sizes between treatment and control conditions were negative across all three dependent measures. This finding suggested that when WMC values were set to a low level, an advantage was found for the control when compared to the treatment conditions. Thus, no support was found for the notion that strategy conditions facilitated compensatory processing for children with low WMC.

For average achievers with high WMC values, ESs of moderate magnitude occurred for strategy conditions when compared to the control condition (ESs ranged from .42 to .57) on the posttest

problem solving measure. However, for average achievers with WMC values in the middle range, no advantages were found for a particular strategy condition when compared to the control condition on posttest measures of problem solving or calculation accuracy (ESs ranged from -1.20 to -1.29). An advantage for average achievers on the posttest operation span measure occurred for those with relatively lower WMC values. The visual-only condition was particularly robust when compared to the control condition ( $ES = 1.89$ ).

## Discussion

This study investigated the role of strategy instruction and working memory capacity on word problem solving accuracy and transfer measures in children with MD. The results showed a significant WMC  $\times$  treatment interaction across all criterion measures. In general, the results indicated that working memory capacity played an important role in moderating the effectiveness of strategies on posttest performance outcomes. For children with MD, positive effect sizes in favor of verbal or visual conditions occurred when compared to the control condition on posttest measures of problem solving and calculation accuracy. However, these effects were isolated to children with relatively higher WMC scores. In contrast, no significant strategy treatment advantages occurred relative to the control condition for those children with MD and relatively low WMC. The treatment condition found particularly advantageous to children with MD who had relatively higher WMC across all dependent measures was the visual-only condition. Children with higher WMC, especially those with MD, were more likely to benefit from the diagramming condition than were those with lower WMC. The results will now be discussed in terms of the question that directed the study: Does WMC play an important role in accounting for cognitive strategy outcomes?

Although an answer to this question is in the affirmative, the results must be placed in the context of the four models discussed in the introduction. One model argued that if reading, computation, and general fluid intelligence were relatively intact (in the normal range) for children with MD (as was the case in this study), then the reliable use of cognitive strategies supersedes the role that any individual differences in WMC might play. For the present study, scores for children with MD in the areas of reading, calculation, and fluid intelligence were in the normal range. Further, performance on these measures was statistically comparable among children with MD across treatment conditions. Thus, one would predict from this model minimal variation in strategy outcomes as a function of WMC. However, I found, in contrast to this hypothesis, that WMC interacted with strategy conditions on all criterion measures. Thus, the results do not support the notion that WMC plays a secondary role in problem-solving outcomes related to treatment conditions for children with MD.

A second model suggested that a limited-capacity WM system underlies word problem solving difficulties in children with MD. This model is consistent with the notions of several theorists, who adopt a general resource approach in which individual differences on cognitive and aptitude measures draw on a limited supply of WM resources (e.g., Colom et al., 2008). This model assumes that although WMC may act in tandem with other processes, this general system may operate independent of strategy conditions. This model was not supported because a significant moderating

effect (interaction) emerged between WMC and strategy conditions, suggesting that certain strategies draw upon more working memory resources than others. As shown in Table 4, across all criterion measures, children with MD who had relatively higher WMC benefited from strategy conditions (i.e., verbal-only, visual-only) relative to the control condition. Such was not the case for children with MD who had low WMC scores.

A third model suggests that strategy training compensates for individual differences in WMC. Some studies have shown that strategy training helps low-span participants allocate WM resources more efficiently than it does high-span participants (e.g., Turley-Ames & Whitfield, 2003). Thus, I expected that children with MD, especially those with relatively lower WM span, would benefit more from strategy instruction when compared to the control condition than would average achieving children (children with high spans). Such was not the case in this study. Overall, this study showed that children with MD and low WMC in the strategy conditions did not improve their problem-solving performance relative to those in the control condition.

Thus, the model I prefer suggested that training in cognitive strategies was more likely to improve problem-solving outcomes for children with MD but with a relatively larger WMC. This is because these children have spare WM sources with which to effectively utilize these strategies. The general patterns of the current study are in line with this model. The results show that WMC, as a continuous variable, interacted with strategy conditions in predicting solution accuracy. There are, however, at least four qualifications to the results.

First, the potential moderating effects of WMC may change with longer intervention periods. Models of skill acquisition (e.g., Ackerman, 1988) suggest that WMC may be important in the early phases of skill acquisition but that it becomes less important with longer interventions, as the implementation of strategies is automatized. Although this study cannot test this hypothesis, it may be the case with repeated use of strategies that the effects of WMC and the disadvantages of strategies in children with MD would be reduced.

Second, adjusted posttest scores as a function of WMC values were estimates from a simple linear regression. It is unlikely that I had enough children with MD performing at a high WMC level or enough average achieving children performing at a low level of WMC to capture subtle differences in treatment outcomes. Thus, my predicted adjusted posttest means (adjusted least square means) may require comparisons at setting WMC to less extreme values. However, when I considered effect sizes computed on adjusted posttest scores that were not conditional on setting WMC to specific levels, high ESs still emerged in favor of the verbal-only and visual-only conditions (1.09, .92, respectively) relative to the control condition for problem-solving accuracy.

Third, the WMC effect for the verbal + visual condition is unclear. Although children with MD who have relatively higher WMC were more likely to benefit from strategy training when compared to those in the control condition, especially for those under visual-only conditions, only small effects were found for the verbal + visual condition regardless of variations in WMC (ESs varied from .02 to .20). I assumed that the verbal + visual condition would increase the child's chances to draw upon separate verbal and visual-spatial storage capacities. Thus, the combination of these storage systems, I assumed, would open up the

possibility that more information could be processed and retained without making excessive demands on WMC (Mayer, 2005). Such did not appear to be the case in this study. It is possible that children with MD may have preferred an activation of a single storage system and may possibly have viewed the combination of verbal (attention to cues) and visual (diagramming) information as distracting or as interfering with more efficient processing. No doubt, further research on this issue is necessary.

Finally, the mechanism that played a role toward improving operation span performance in both ability groups is also unclear. One possible explanation was that the operation span measure was a novel measure and the strategy conditions may have provided some practice in working memory. That is, participants were provided practice in recalling targeted information in the context of distracting information (identifying relevant and irrelevant propositions within word problems), a process attributed to working memory (e.g., Engle et al., 1999). This explanation is consistent with studies that have attempted to directly intervene on working memory performance and influence achievement (e.g., Holmes, Gathercole, & Dunning, 2009; Klingberg et al., 2005). No studies that I am aware of, however, have shown that strategy training within an academic domain (word problem solving) directly influences WM or vice versa (e.g., see Holmes et al., 2009, for discussion of the sleeper effect). Perhaps the approach I took to enhance transfer by embedding working memory demands (load) within the curriculum may be an important avenue in future research. It may also be the simple case, however, that because basic calculation was involved in the training, and because calculation was embedded in the operation span measure, this may have accounted for the transfer effects.

## Implications

Our findings have several applications to current research. First, the study may account for why some children benefit from strategy instructions and others do not. I found that a key variable in accounting for the outcomes was WMC. Clearly, WMC would not be the only variable across studies to account for the outcomes; however, the role of WMC in this study appeared to be fairly robust. It may be the case that when children with computation and/or reading difficulties are included in the analysis that effects would be different. Thus, despite the poor treatment outcomes for children with MD and with low WMC relative to the control condition, it is important to note that children in this study had reading and computation scores within the average range.

Second, for children with MD and low WMC, none of the strategies were found particularly effective relative to the control condition. In fact, the visual-spatial strategy condition, which included diagramming, yielded substantially lower posttest scores than did the control condition on posttest measures of problem solving ( $ES = -1.07$ ) and calculation ( $ES = -.98$ ). This finding aligns with several studies that have suggested that visual-spatial WM (represented by the visual-spatial sketchpad) is closely linked with MD (e.g., Bull, Espy, & Wiebe, 2008). However, a recent meta-analysis synthesizing research on cognitive studies of MD (Swanson & Jerman, 2006) suggests that memory deficits are more apparent in the verbal than the visual spatial WM domain. My findings do suggest, however, that a moderate advantage was found relative to the control condition by combining verbal and



visual training for MD children with WMC values in the low range ( $ES = .14$ ) and the relatively high range ( $ES = .20$ ). The suggestion, perhaps, is that both routes are important route for remediation.

An obvious question emerges as to why the visual-spatial strategy (diagramming) alone condition favored children with higher WMC value scores compared to those with low WMC scores. My best explanation is that the use of diagrams is resource demanding. It is also possible that not all children had adequate resources to enact this visual strategy without placing excessive demands on working memory. The visual-spatial strategy, however, may have provided a technique that allowed children with high WMC to focus on the relevant aspects of the task. Diagramming numbers might have activated the relevant information while preventing irrelevant information from interfering with problem-solving solutions. Taken together, the results suggest that visual diagramming is an effective intervention for some children with MD in order to increase solution accuracy.

Third, verbal-only and visual-only strategy conditions facilitated calculation proficiency for children with MD and relatively higher WMC relative to the control condition, but they decreased performance for average achievers (see Table 4). Improvement in calculation was part of each lesson plan, and practice and feedback therefore could have played a role in the performance of children with MD. As shown in the standard scores reported in Table 1, children with MD had lower calculation scores than did average achieving children; therefore, strategy instruction may have provided an additional boost in performance. The outcomes for the average achievers, however, are less clear. I infer that the outcomes may be related to classroom testing. In all classrooms, children were exposed to daily 1-minute calculation tests (part of a curriculum based assessment measure); therefore, I infer that as average achieving children became increasingly fluent in calculations, strategy conditions may have actually interfered (i.e., slowed them down) with skills that were fairly well automatized.

A final application relates to improvement on a norm-referenced test. The majority of intervention studies for problem solving have shown gains on experimental measures and less gain on standardized measures (for reviews, see Jitendra & Xin, 1997; Powell, 2011). Thus, in the current study, I was able to improve performance substantially on materials related to standardized tests. Although I used  $z$  scores (based on raw scores) rather than national norms to compare treatment conditions, it is important to note that in the treatment  $\times$  covariate analyses, estimated adjusted problem solving posttest scores for children with MD and relative higher WMC exceeded scores for those in the control condition, especially those in the verbal and/or visual condition.

## Summary

Taken together, these findings suggest that WMC moderates the influence of cognitive strategies. The results suggest that solution accuracy for children with MD, relative to the control condition, improved substantially as a function of both verbal and visual strategy training for those with relatively higher WMC. Additionally, weak support was found for the assumption that strategy conditions compensate for low WMC in children with MD.

## References

- Ackerman, P. L. (1988). Determinants of individual differences in skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General*, 117, 288–318. doi:10.1037/0096-3445.117.3.288
- Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solving. *Psychological Review*, 94, 192–210. doi:10.1037/0033-295X.94.2.192
- Andersson, U. (2010). Skill development in different components of arithmetic and basic cognitive functions: Findings from a 3-year longitudinal study of children with different types of learning difficulties. *Journal of Educational Psychology*, 102, 115–134. doi:10.1037/a0016838
- Baddeley, A. D., & Logie, R. H. (1999). The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28–61). Cambridge, United Kingdom: Cambridge University Press.
- Baker, S., Gersten, R., & Lee, D. (2002). A synthesis of empirical research on teaching mathematics to low-achieving students. *Elementary School Journal*, 103, 51–73. doi:10.1086/499715
- Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research*, 40, 373–400. doi:10.1207/s15327906mbr4003\_5
- Brown, V. L., Cronin, M. E., & McEntire, E. (1994). *Test of Mathematical Ability*. Austin, TX: PRO-ED.
- Brown, V. L., Hammill, D., & Weiderholt, L. (1995). *Test of Reading Comprehension*. Austin, TX: PRO-ED.
- Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. *Developmental Neuropsychology*, 33, 205–228. doi:10.1080/87565640801982312
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Colom, R., Abad, F. J., Quiroga, M. Á., Shih, P. C., & Flores-Mendoza, C. (2008). Working memory and intelligence are highly related constructs, but why? *Intelligence*, 36, 584–606. doi:10.1016/j.intell.2008.01.002
- Connolly, A. J. (1998). *KeyMath-Revised/Normative Update*. Circle Pines, MN: American Guidance Service.
- Engle, R. W., Cantor, J., & Carullo, J. J. (1992). Individual differences in working memory and comprehension: A test of four hypotheses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 972–992. doi:10.1037/0278-7393.18.5.972
- Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent variable approach. *Journal of Experimental Psychology: General*, 128, 309–331. doi:10.1037/0096-3445.128.3.309
- Fletcher, J. M., Espy, K. A., Francis, P. J., Davidson, K. C., Rourke, B. P., & Shaywitz, S. E. (1989). Comparisons of cutoff and regression-based definitions of reading disabilities. *Journal of Learning Disabilities*, 22, 334–338. doi:10.1177/002221948902200603
- Fuchs, L. S., Fuchs, D., Prentice, K., Burch, M., Hamlett, C. L., Owen, R., Schroeter, K. (2003). Enhancing third-grade students' mathematical problem solving with self-regulated learning strategies. *Journal of Educational Psychology*, 95, 306–315. doi:10.1037/0022-0663.95.2.306
- Fuchs, L. S., Fuchs, D., Prentice, K., Hamlett, C. L., Finelli, R., & Courey, S. J. (2004). Enhancing mathematical problem solving among third-grade students with schema-based instruction. *Journal of Educational Psychology*, 96, 635–647. doi:10.1037/0022-0663.96.4.635
- Fuchs, L. S., Fuchs, D., Stuebing, K., Fletcher, J. M., Hamlett, C. L., & Lambert, W. (2008). Problem solving and computational skill: Are they shared or distinct aspects of mathematical cognition? *Journal of Educational Psychology*, 100, 30–47. doi:10.1037/0022-0663.100.1.30
- Fuchs, L. S., Zumeta, R. O., Schumacher, R. F., Powell, S. R., Seethaler, P. M., Hamlett, C. L., & Fuchs, D. (2010). The effects of schema-broadening instruction on second graders' word-problem performance

- and their ability to represent word problems with algebraic equations: A randomized control study. *Elementary School Journal*, 110, 446–463. doi:10.1086/651191
- Geary, D. C. (2003). Math disabilities. In H. L. Swanson, K. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 199–212). New York, NY: Guilford Press.
- Geary, D. C. (2010). Mathematical disabilities: Reflections on cognitive, neuropsychological, and genetic components. *Learning and Individual Differences*, 20, 130–133. doi:10.1016/j.lindif.2009.10.008
- Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research*, 79, 1202–1242. doi:10.3102/0034654309334431
- Holmes, J., Gathercole, S. E., & Dunning, D. L. (2009). Adaptive training leads to sustained enhancement of poor working memory in children. *Developmental Science*, 12, 9–15. doi:10.1111/j.1467-7687.2009.00848.x
- Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.
- Hresko, W., Schlieve, P. L., Herron, S. R., Swain, C., & Sherbenou, R. (2003). *Comprehensive Math Abilities Test*. Austin, TX: PRO-ED.
- Jitendra, A. K., Griffin, C. C., McGoey, K., Gardill, M. C., Bhat, P., & Riley, T. (1998). Effects of mathematical word problem solving by students at risk for mild disabilities. *Journal of Educational Research*, 91, 345–355. doi:10.1080/00220679809597564
- Jitendra, A., & Xin, Y. P. (1997). Mathematical word-problem-solving instruction for students with mild disabilities and students at risk for math failure: A research synthesis. *Journal of Special Education*, 30, 412–438. doi:10.1177/002246699703000404
- Judd, C. M., McClelland, G. H., & Smith, E. R. (1996). Testing treatment by covariate interactions when treatment varies within subjects. *Psychological Methods*, 1, 366–378. doi:10.1037/1082-989X.1.4.366
- Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., . . . Westerberg, H. (2005). Computerized training of working memory in children with ADHD: A randomized, controlled trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, 44, 177–186. doi:10.1097/00004583-200502000-00010
- Kolloffel, B., Eysink, T., de Jong, T., & Wilhelm, P. (2009). The effects of representation format on learning combinatorics from an interactive computer. *Instructional Science*, 37, 503–517. doi:10.1007/s11251-008-9056-7
- Lazer, A. A., & Zerbe, G. O. (2011). Solutions for determining the significant region using the Johnson–Neyman type procedure in generalized linear (mixed) models. *Journal of Educational and Behavioral Statistics*, 36, 699–719. doi:10.3102/107699861039889
- Lee, K., Ng, S.-F., Ng, E.-L., & Lim, Z.-Y. (2004). Working memory and literacy as predictors of performance on algebraic word problems. *Journal of Experimental Child Psychology*, 89, 140–158. doi:10.1016/j.jecp.2004.07.001
- Leon, A. C., Portera, L., Lowell, M. A., & Rheinheimer, D. (1998). A strategy to evaluate covariance by group interaction in an analysis of variance. *Psychopharmacology Bulletin*, 34, 805–809.
- Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2010). *SAS for mixed models*. (2nd ed.). Cary, NC: SAS.
- Looi, C.-K., & Lim, K.-S. (2009). From bar diagrams to letter-symbolic algebra: A technology-enabled bridging. *Journal of Computer Assisted Learning*, 25, 358–374. doi:10.1111/j.1365-2729.2009.00313.x
- Mastropieri, M. A., Scruggs, T. E., & Shiah, R.-L. (1997). Can computers teach problem-solving strategies to children with mild mental retardation? *Remedial and Special Education*, 18, 157–164. doi:10.1177/074193259701800304
- Mayer, R. E. (2005). Cognitive theory and multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31–48). New York, NY: Cambridge University Press.
- Mayer, R. E., & Hegarty, M. (1996). The process of understanding mathematical problem solving. In R. J. Sternberg & T. Ben-Zeev (Eds.), *The nature of mathematical thinking* (pp. 29–54). Mahwah, NJ: Erlbaum.
- Mazzocco, M. M. M., Devlin, K. T., & McKenney, S. J. (2008). Is it a fact? Timed arithmetic performance of children with mathematical learning disabilities (MLD) varies as a function of how MLD is defined. *Developmental Neuropsychology*, 33, 318–344. doi:10.1080/87565640801982403
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., & Howerter, A. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, 41, 49–100. doi:10.1006/cogp.1999.0734
- Montague, M. (2008). Self-regulation strategies to improve mathematical problem solving for students with learning disabilities. *Learning Disability Quarterly*, 31, 37–44. doi:10.2307/30035524
- Montague, M., Warger, C., & Morgan, T. H. (2000). Solve It! strategy instruction to improve mathematical problem solving. *Learning Disabilities Research & Practice*, 15, 110–116. doi:10.1207/SLDRP1502\_7
- Morris, N., & Jones, D. M. (1990). Memory updating in working memory: The role of central executive. *British Journal of Psychology*, 81, 111–121. doi:10.1111/j.2044-8295.1990.tb02349.x
- Ng, S. F., & Lee, K. (2009). The model method: Singapore children’s tool for representing and solving algebraic word problems. *Journal for Research in Mathematics Education*, 40(3), 282–313.
- Passolunghi, M. C., Cornoldi, C., & De Liberto, S. (1999). Working memory and intrusions of irrelevant information in a group of specific poor problem solvers. *Memory & Cognition*, 27, 779–790. doi:10.3758/BF03198531
- Passolunghi, M. C., & Siegel, L. S. (2001). Short-term memory, working memory, and inhibitory control in children with difficulties in arithmetic problem solving. *Journal of Experimental Child Psychology*, 80, 44–57. doi:10.1006/jecp.2000.2626
- Pearson Publishers. (2009). *Scott Forsman-Addison Wesley EnVisionMath*. New York, NY: Author.
- Powell, S. R. (2011). Solving word problems using schemas: A review of literature. *Learning Disabilities Research & Practice*, 26, 94–108. doi:10.1111/j.1540-5826.2011.00329.x
- Psychological Corporation. (1992). *Wechsler Individual Achievement Test*. San Antonio, TX: Harcourt Brace.
- Raven, J. C. (1976). *Colored progressive matrices test*. London, England: Lewis.
- Resendes, M., & Azin, M. (2008). *A study of the effects of Pearson’s 2009 enVisionMATH Program*. Jackson WY: PRES Associates.
- Rogosa, D. (1980). Comparing nonparallel regression lines. *Psychological Bulletin*, 88, 307–321. doi:10.1037/0033-2909.88.2.307
- Siegel, L. S., & Ryan, E. B. (1989). The development of working memory in normally achieving and subtypes of learning disabled. *Child Development*, 60, 973–980. doi:10.2307/1131037
- Stock, P., Desoete, A., & Roeyers, H. (2010). Detecting children with arithmetic disabilities from kindergarten: Evidence from a 3-year longitudinal study on the role of preparatory arithmetic abilities. *Journal of Learning Disabilities*, 43, 250–268. doi:10.1177/0022219409345011
- Swanson, H. L. (1992). Generality and modification of working memory among skilled and less skilled readers. *Journal of Educational Psychology*, 84, 473–488. doi:10.1037/0022-0663.84.4.473
- Swanson, H. L. (1995). *S-Cognitive Processing Test (S-CPT): A dynamic assessment measure*. Austin, TX: PRO-ED.
- Swanson, H. L., & Beebe-Frankenberger, M. (2004). The relationship between working memory and mathematical problem solving in children at risk and not a risk for serious math difficulties. *Journal of Educational Psychology*, 96, 471–491. doi:10.1037/0022-0663.96.3.471



- Swanson, H. L., Cooney, J. B., & Brock, S. (1993). The influence of working memory and classification ability on children's word problem solution. *Journal of Experimental Child Psychology*, 55, 374–395. doi/10.1006/jecp.1993.1021
- Swanson, H. L., & Jerman, O. (2006). Math disabilities: A selective meta-analysis of the literature. *Review of Educational Research*, 76, 249–274. doi:10.3102/00346543076002249
- Swanson, H. L., Jerman, O., & Zheng, X. (2008). Growth in working memory and mathematical problem solving in children at risk and not at risk for serious math difficulties. *Journal of Educational Psychology*, 100, 343–379. doi:10.1037/0022-0663.100.2.343
- Swanson, H. L., Kehler, P., & Jerman, O. (2010). Working memory, strategy knowledge, and strategy instruction in children with reading disabilities. *Journal of Learning Disabilities*, 43, 24–47. doi:10.1177/0022219409338743
- Sweller, J. (1988). Cognitive load during problem solving: Effects of learning. *Cognitive Science*, 12, 257–285. doi:10.1016/0364-0213(88)90023-7
- Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 19–30). New York, NY: Cambridge University Press.
- Turley-Ames, K. J., & Whitfield, M. (2003). Strategy training and working memory performance. *Journal of Memory and Language*, 49, 446–468. doi:10.1016/S0749-596X(03)00095-0
- van Garderen, D. (2007). Teaching students with LD to use diagrams to solve mathematical word problems. *Journal of Learning Disabilities*, 40, 540–553. doi:10.1177/00222194070400060501
- What Works Clearinghouse. (2006). *Technical details of WWC-conducted comparisons (9–12-2006)*. Retrieved January 1, 2010, from <http://ies.ed.gov/ncee/WWC>
- Widaman, K. F. (2006). Best practices in quantitative methods for developmentalists: III. Missing data: What to do with or without them. *Monographs of the Society for Research in Child Development*, 71(3), 42–64. doi:10.1111/j.1540-5834.2006.00404.x
- Wilkinson, G. S. (1993). *The Wide Range Achievement Test*. Wilmington, DE: Wide Range Inc.
- Xin, Y. P. (2008). The effect of schema-based instruction in solving mathematics word problems: An emphasis on prealgebraic conceptualization of multiplicative relations. *Journal for Research in Mathematics Education*, 39, 526–551.
- Xin, Y. P., & Jitendra, A. K. (1999). The effects of instruction in solving mathematical word problems for students with learning problems: A meta-analysis. *Journal of Special Education*, 32, 207–225. doi:10.1177/002246699903200402

Received April 4, 2012

Revision received October 25, 2013

Accepted December 15, 2013 ■

# Learning With Retrieval-Based Concept Mapping

Janell R. Blunt and Jeffrey D. Karpicke  
Purdue University

Students typically create concept maps while they view the material they are trying to learn. In these circumstances, concept mapping serves as an elaborative study activity—students are not required to retrieve the material they are learning. In 2 experiments, we examined the effectiveness of concept mapping when it is used as a retrieval practice activity. In Experiment 1, students read educational texts and practiced retrieval either by writing down as many ideas as they could recall in paragraph format or by creating a concept map (retrieval-based concept mapping). In Experiment 2, we factorially crossed the format of the activity (paragraph vs. concept map) and the presence or absence of the text (i.e., whether the activity involved repeated studying or retrieval practice). On a final test 1 week later that assessed verbatim knowledge and inferencing, both paragraph and concept map retrieval practice formats produced better performance than additional studying, but the 2 retrieval formats themselves did not differ. The results demonstrate the effectiveness of concept mapping when it is used as a retrieval practice activity and show that retrieval itself, rather than merely the act of writing, drives the benefits of retrieval-based learning activities.

**Keywords:** retrieval practice, concept mapping, learning, writing, study strategies

Learning is often viewed as a process that occurs primarily when people encode or study new material, and the best learning is thought to occur when students elaborate on what they are studying by forming meaningful connections and creating enriched knowledge structures. Retrieval, which occurs when students take tests, is viewed as an assessment of learning that occurred in prior study experiences but is not thought to create learning itself. In contrast to the latter assumption, a great deal of recent research has shown that practicing retrieval creates long-term, meaningful learning, sometimes even more learning than elaborative encoding activities (see Karpicke, 2012; Karpicke & Blunt, 2011). Our purpose in this article was to examine the effectiveness of two different retrieval practice formats: retrieving by writing information in paragraph format (a common way to induce retrieval practice; Roediger & Karpicke, 2006), and retrieving by creating what we refer to as *retrieval-based concept maps*.

The exact mechanisms underlying the effects of retrieval practice have not yet been specified, but the idea that retrieval practice

effects stem from elaborative study processes has recently been called into question. If elaboration were responsible for retrieval practice effects, then engaging in repeated retrieval should produce the same or similar effects as engaging in repeated elaborative studying. Recent research, however, has shown that repeated retrieval consistently produces greater levels of long-term learning than elaborative studying. For example, Karpicke and Smith (2012) found that retrieval practice produced superior long-term retention relative to imagery-based and verbal elaborative study methods (see too Karpicke & Blunt, 2011). Instead of elaborative study processes, the benefits of retrieval practice are currently thought to stem from processes involved recollecting the context of a prior learning episode (Karpicke, Lehman, & Aue, in press). Remembering what occurred at a particular place and time is not necessary during a semantic elaboration task, but it is inherent to retrieval practice. As evidence for this account of retrieval-based learning, Karpicke and Zaromb (2010) showed that having people intentionally retrieve a prior event (in their experiments, the completions to word fragments) led to greater subsequent retention relative to asking people to generate knowledge without thinking back to the past (e.g., completing fragments with the first words that came to mind). Therefore, an essential component of retrieval-based learning is what Tulving (1983) called being in an *episodic retrieval mode*, which refers to the act of thinking back to what occurred in a particular place and time in the past. Retrieval-based learning activities should be aimed at guiding students to intentionally recollect prior experiences.

In the experiments reported here, we examined the effectiveness of using one popular learning task, *concept mapping*, as a retrieval-based learning activity. Concept mapping is a graphic organizational technique in which students create node-and-link diagrams, where nodes represent concepts and links connecting the nodes represent relations among the concepts (see Figure 1; Novak & Gowin, 1984). Typically, students construct concept maps while

---

This article was published Online First February 17, 2014.

Janell R. Blunt and Jeffrey D. Karpicke, Department of Psychological Sciences, Purdue University.

This research was supported in part by grants from the National Science Foundation (DUE-0941170 and DRL-1149363) and the Institute of Education Sciences in the U.S. Department of Education (R305A110903). Janell R. Blunt is supported by National Science Foundation Graduate Research Fellowship 2012124747. The opinions expressed are those of the authors and do not represent the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

We thank Breanne Lawler, Stephanie Angel, Kyle Ward, Mindi Cogdill, and Brittany Etchison for helping collecting the data and Philip Grimaldi for help with computer programming.

Correspondence concerning this article should be addressed to Janell R. Blunt, Department of Psychological Science, Purdue University, 703 Third Street, West Lafayette, IN 47907-2081. E-mail: jrblunt@purdue.edu



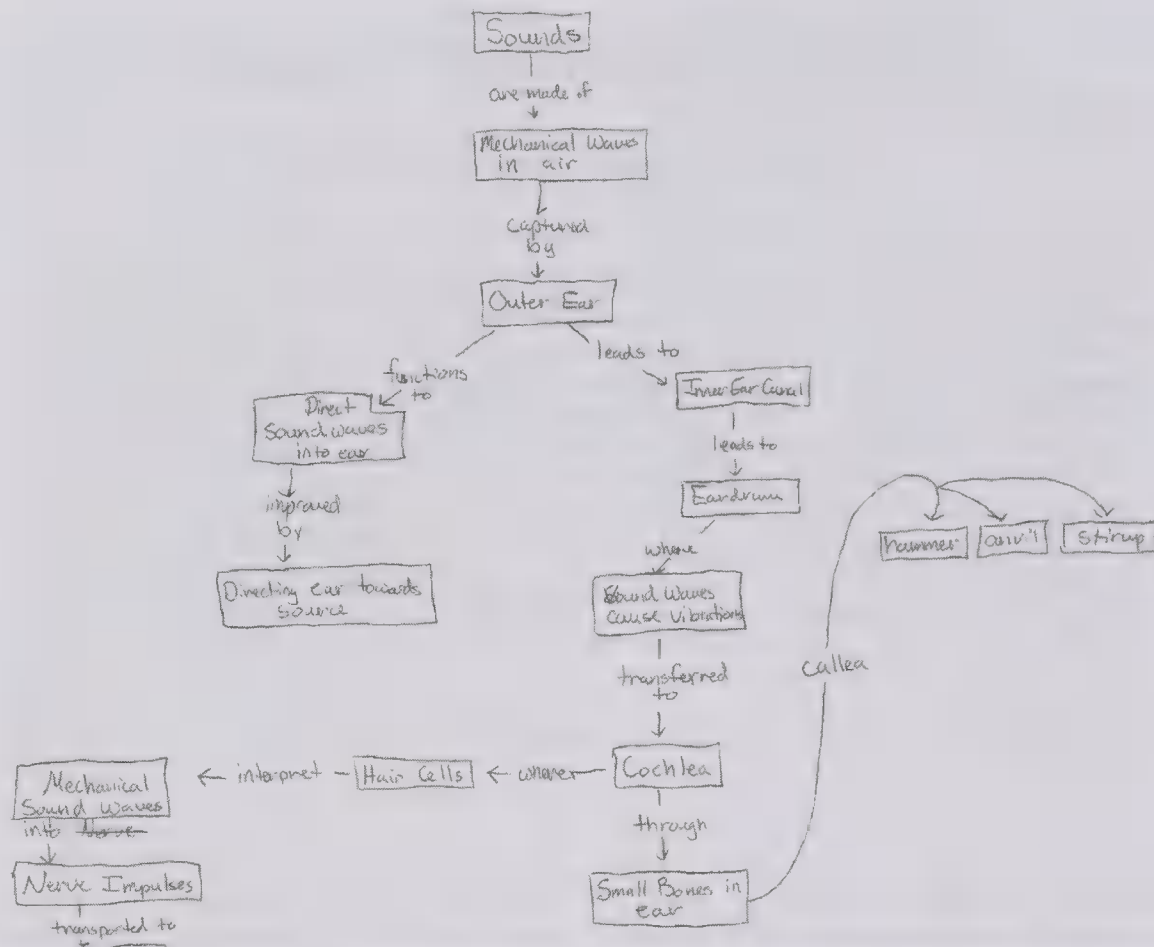


Figure 1. An example of a concept map created by a student in Experiment 1.

they view the materials they are learning. Although this presumably allows students to enrich the material by encoding meaningful relationships among concepts, when students create concept maps while viewing the to-be-learned materials, they are not required to practice retrieving the materials.

Recently, we carried out two experiments in which we directly compared the effectiveness of retrieval practice and elaborative studying with concept mapping (Karpicke & Blunt, 2011). Students studied educational texts on various science topics and either practiced retrieval or created concept maps of the texts. In the concept map conditions, students created concept maps while viewing the texts, whereas in the retrieval practice conditions, students read the texts and practiced retrieval by writing down as much of the material as they could recall without viewing the texts (a standard way of implementing retrieval practice for educational texts, which we refer to in this article as *paragraph format*; see Karpicke & Roediger, 2010; Roediger & Karpicke, 2006). The effects of these activities were assessed on final tests 1 week after the original learning phase. Practicing retrieval produced better long-term learning than elaborative concept mapping on final short-answer questions that assessed verbatim knowledge (items stated directly in the original text) and inferential knowledge (questions that required students to connect multiple concepts in the text). Furthermore, the benefits of retrieval practice were observed not only on short-answer questions but also on final assessments that involved creating a concept map of the material (Experiment 2 in Karpicke & Blunt, 2011). Thus, practicing retrieval produced more learning than creating concept maps when

the concept mapping activity was used as an elaborative study method.

Concept mapping could be used as technique to implement retrieval practice, and there are reasons to expect that concept mapping might serve as an effective retrieval-based learning activity. Specifically, concept mapping requires students to identify the main concepts in a text (Hay, Kinchin, & Lygo-Baker, 2008; Stewart, Van Kirk, & Rowell, 1979) and then identify how the concepts are related to each other, which helps focus students on the organizational structure of the material (Vanides, Yin, Tomita, & Ruiz-Primo, 2005). It is also assumed that creating concept maps helps student use their own prior knowledge to identify how concepts might be related (Novak, 1976). Thus, concept mapping is thought to promote not only students' verbatim knowledge and comprehension but also students' abilities to make inferences about what they are learning (Novak & Gowin, 1984).

Alternatively, there are also reasons to expect that concept mapping might not serve as an effective retrieval-based learning activity. When students freely recall material, they must adopt a retrieval strategy to guide their recall output (e.g., when recalling texts in paragraph format, students tend to recall in serial order, presumably to preserve the text structure; see Karpicke & Roediger, 2010). Concept mapping might require students to adopt an ineffective retrieval strategy or might disrupt students' default strategies, which could weaken the benefits of retrieval practice. It is also possible that asking students to retrieve knowledge in concept map format could introduce additional cognitive load during the process of retrieval, or the mapping task might function

as a secondary task that divides students' attention. Either factor could reduce the effectiveness of retrieval practice. Finally, retrieval-based concept mapping might produce learning that is equivalent to the learning afforded by practicing retrieval in paragraph format. This outcome would be expected if the organizational processing thought to occur during concept mapping were redundant with relational processing already afforded by paragraph recall and if both retrieval formats effectively allowed students to recollect the prior episodic context, which is the mechanism considered central to retrieval-based learning (Karpicke et al., in press; Karpicke & Zaromb, 2010).

In the present experiments, students read brief educational texts and practiced retrieval by writing in paragraph format or by creating concept maps. The effects of these retrieval practice activities were examined on a delayed short-answer test 1 week after the original learning phase. We also examined students' subjective experiences of the different activity formats. We were especially interested in students' judgments of learning (their predictions of how well they would perform in the future), but we also examined students' ratings of how interesting, difficult, and enjoyable the activities were. The inclusion of these metacognitive judgments allowed us to examine the correspondence between students' actual learning and their predicted performance, which is especially important to examine in light of claims that concept mapping represents "the most important metacognitive tool in science education" (Mintzes, Wandersee, & Novak, 1997, p. 424).

## Experiment 1

Experiment 1 was a conceptual replication of Karpicke and Blunt (2011) with one important change: Rather than having students create concept maps while viewing texts, we had them create concept maps in the absence of the texts. Thus, we directly compared two different retrieval practice formats: concept mapping and paragraph recall. Students read and practiced retrieval of brief science texts. During concept map retrieval practice, students retrieved the material by creating a concept map, whereas during paragraph retrieval practice, students wrote as much of the material as they could recall in paragraph format. The students then made a series of metacognitive judgments (judgments of learning, interest, difficulty, and enjoyment). The effects of the two retrieval practice formats were assessed on a final test 1 week after the original learning phase that included both verbatim and inference short-answer questions.

## Method

**Subjects.** Thirty-two Purdue University undergraduates participated in partial fulfillment of course requirements.

**Materials.** Two brief texts were selected from Cook and Mayer (1988, as described and used by Karpicke & Blunt, 2011). One text, "The Human Ear," had a sequential structure (Meyer, 1975), which means the text described a connected series of events and steps in a process (the sequence of events involved in the process of hearing). The other text, "Make-Up of Human Blood," had an enumeration structure, which means that the text listed and described a series of concepts (the properties of different blood components). The texts were 259 and 236 words in length, respectively.

**Design.** The two retrieval formats (concept map vs. paragraph) were manipulated within subject. Each student studied two texts and practiced retrieval of one text in concept map format and the other in paragraph format. The order of the two texts and the order in which students performed the two learning activities were counterbalanced across students.

**Procedure.** Students were tested in small groups in two sessions. During the learning phase (Session 1), students read one text for 5 min, recalled it for 10 min, reread it for 5 min, and recalled it again for 10 min in one of the two retrieval practice conditions. They then repeated this procedure for the other text and other retrieval practice condition.

Before completing the concept mapping retrieval practice condition, the students were instructed about the nature of the concept mapping activity. They were told that a concept map is a diagram in which concepts are represented as nodes that are linked together with words and phrases. The students were shown an example of a concept map selected from Novak (2005). Then, during recall periods, they were given a sheet of paper and told to recall the text by creating a concept map. Students were allowed to refer to the example concept map, but not to the text, throughout each 10-min recall period. Pilot testing showed that this was enough time for students to reach asymptotic levels of recall under these conditions.

In the paragraph retrieval practice condition, students saw a response box on a computer screen and were told to recall as much of the information from the text as they could by typing their responses on the computer during each 10-min recall period (see Karpicke & Roediger, 2010). Overall, the total amount of learning time was identical in the elaborative concept mapping and retrieval practice conditions.

At the end of each learning activity, the students were asked to predict how much of the material they would remember in 1 week (an aggregate judgment of learning) and to rate the enjoyment, difficulty, and interestingness of the activities. Students made their ratings on a scale from 0% to 100% in increments of 10 (0, 10, 20, . . . 80, 90, 100). At end of Session 1, after completing both activities, students indicated which retrieval practice format they preferred.

The students were dismissed and returned to the laboratory 1 week later for the final short-answer test, which included 10 verbatim questions and four inference questions per text. Examples of questions are shown in the Appendix. During the final test, each question remained on the screen for at least 20 s; at that time, a button labeled "Next" appeared on the screen, and students pressed the button to proceed to the next question. Students were encouraged to take as much time as needed to answer the questions. At the end of the second session, the students were debriefed and thanked for their participation.

## Results

An initial analysis indicated that there were no differences among the counterbalancing orders, so the results have been collapsed across orders. There was a difference between texts such that initial and final performance was better on the "Make-Up of Human Blood" text than on the "Human Ear" text. However, text did not interact with any other factors in the experiment, so the results have been collapsed across texts.



**Scoring.** The texts were divided into 30 idea units for scoring purposes. Both the paragraph and concept map protocols were scored using the same criteria: Students were given 1 point for each idea unit recalled (Karpicke & Blunt, 2011; Karpicke & Roediger, 2010). On the final short-answer test, correct responses were given 1 point, and partially correct responses were given partial credit (e.g., .75, .50, or .25 points, depending on completeness of the response). Two independent raters scored all recall protocols and short-answer tests, and a third rater resolved all discrepancies to reach 100% agreement.

**Learning performance.** The left side of Table 1 shows performance during the learning phase in Experiment 1 (the proportion of idea units recalled in each condition). Collapsed across retrieval formats, the proportion of ideas recalled increased from Period 1 to Period 2 (.39 vs. .55),  $t(31) = 10.34$ ,  $d = 1.83$ , 95% confidence interval (CI) [1.25, 2.39].<sup>1</sup> Students recalled more ideas in paragraph format than in concept map format. This pattern occurred in Period 1,  $t(31) = 3.77$ ,  $d = 0.66$ , 95% CI [0.28, 1.04], and in Period 2,  $t(31) = 5.52$ ,  $d = 0.98$ , 95% CI [0.55, 1.39]. We examined the differences in initial recall in the concept map and paragraph conditions in a post hoc analysis reported in a later section.

**Final short-answer performance.** Figure 2 shows performance on the final short-answer test that occurred 1 week after the initial learning phase. Performance was essentially equivalent in the concept map and paragraph retrieval practice format conditions. There were only small differences, slightly favoring the paragraph format over the concept map format, on the verbatim questions (.68 vs. .62),  $t(31) = 1.07$ ,  $d = 0.19$ , 95% CI [−0.16, 0.54], and on the inference questions (.84 vs. .82),  $t(31) = 0.41$ ,  $d = 0.07$ , 95% CI [−0.28, 0.42].

**Subjective ratings.** The right panel of Figure 2 shows students’ judgments of learning, and Table 2 shows students’ additional ratings of their experiences during the learning tasks. There were very small differences in students’ judgments of learning,  $t(31) = 0.26$ ,  $d = 0.05$ , 95% CI [−0.30, 0.39]; ratings of enjoyment,  $t(31) = 0.31$ ,  $d = 0.05$ , 95% CI [−0.29, 0.40]; ratings of task difficulty,  $t(31) = 0.33$ ,  $d = 0.06$ , 95% CI [−0.29, 0.40]; and ratings of the interestingness of the tasks,  $t(31) = 1.04$ ,  $d = 0.18$ , 95% CI [−0.17, 0.53]. However, at the end of the initial learning phase, when students were asked to indicate which format they preferred, the majority of students preferred the paragraph format (20/32 students = 63%) to the concept map format (12/32 students = 37%).

Table 1  
*Proportion of Idea Units Produced in Each Learning Period in Experiments 1 and 2*

Learning activity	Experiment 1		Experiment 2	
	Period 1	Period 2	Period 1	Period 2
Retrieval practice (no text)				
Concept map	.33 (.02)	.45 (.03)	.24 (.03)	.39 (.03)
Paragraph	.44 (.03)	.64 (.03)	.27 (.03)	.48 (.04)
Repeated study (text)				
Concept map	—	—	.48 (.02)	.58 (.03)
Paragraph	—	—	.53 (.04)	.62 (.03)

Note. Standard errors of the means are shown in parentheses.

**Conditional analysis.** In the next two sections, we report two sets of analyses aimed at exploring differences in recall in the concept map and paragraph conditions. The left portion of Table 3 shows the results of an analysis of the relationship between initial learning performance and final short-answer performance (collapsed across question type) in Experiment 1. In order to analyze the fate of idea units on the final test, we coded short-answer questions based on the idea unit or units required to answer the questions. Verbatim questions typically required access to a single idea unit (collapsed across texts,  $M = 1.3$  idea units per verbatim question). For example, the question “What happens when hemoglobin combines with oxygen?” corresponded to the idea unit “Hemoglobin releases oxygen to the lungs.” Inference questions required access to multiple idea units (collapsed across texts,  $M = 2.3$  idea units per inference question). For example, the question “What would happen if blood did not contain white blood cells, and bacteria were introduced to the body?” relies on the following idea units: (a) “White blood cells are mainly disease fighters”; (b) “White blood cells digest bacteria and other foreign material”; and (c) “When there is an infection somewhere in the body, white blood cells move toward it.”

We followed Tulving’s (1964) method to analyze the correspondence in recall of individual idea units across two tests (see also Karpicke & Zaromb, 2010).  $C_1$  refers to idea units produced in either Period 1 or 2 in the initial learning phase, and  $N_1$  refers to idea units that were not produced in the initial learning phase.  $C_2$  refers to short-answer questions correctly answered on the final short-answer test, and  $N_2$  refers to questions not correctly answered on the final test. As shown in Table 3, the joint probability of recalling an idea initially and correctly answering a final short-answer question ( $C_1C_2$ ) was greater in the paragraph condition than in the concept map condition,  $t(31) = 3.68$ ,  $d = 0.65$ , 95% CI [0.26, 1.03]. Likewise, the probability of not recalling an idea but then correctly answering a final question ( $N_1C_2$ ) was greater in the concept map condition than in the paragraph condition,  $t(31) = 2.45$ ,  $d = 0.43$ , 95% CI [0.07, 0.79]. Together, these results reflect the fact that students initially recalled more ideas in paragraph format than they did in concept map format, yet the conditions produced equivalent levels of final short-answer performance. (We explore this pattern further in the analysis reported in the next section.) There was a small difference in *intertest forgetting* (the probability of recalling an idea but then failing to answer a short-answer question;  $C_1N_2$ ) across conditions, with the paragraph condition showing slightly less forgetting,  $t(31) = 1.67$ ,  $d = 0.30$ , 95% CI [−0.06, 0.65]. Finally, proportion of ideas not recalled or expressed on either test ( $N_1N_2$ ) was greater in the concept map condition relative to the paragraph condition,  $t(31) = 2.32$ ,  $d = 0.41$ , 95% CI [0.05, 0.77].

**Initial recall and normative importance.** Students might have produced fewer ideas during initial concept map recall relative to initial paragraph recall because they adopted different output strategies in the two tasks. We reasoned that students might selectively produce only “important” ideas under concept map

<sup>1</sup> We report standardized mean differences ( $ds$ ) and 95% confidence intervals around the effect size estimates (see Cumming, 2012), which were calculated using the Methods for the Behavioral, Educational, and Social Sciences (MBESS) package for R (Kelley, 2007).

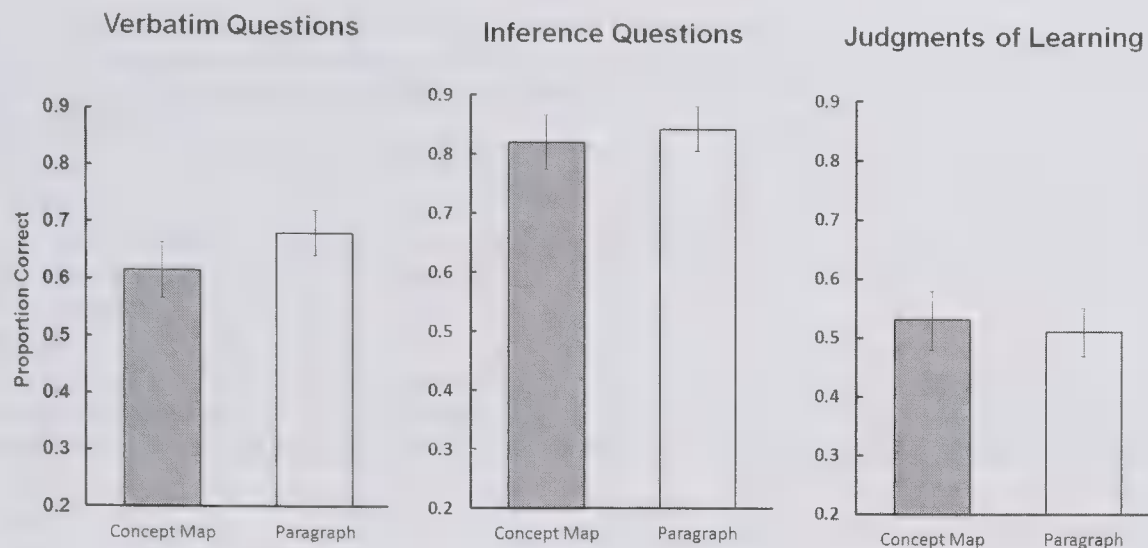


Figure 2. Final short-answer performance for verbatim questions and inference questions (left and middle panels), and judgments of learning (right panel) in Experiment 1. Error bars represent standard errors of the means.

conditions. To examine this possibility, we had 16 undergraduate students, who were not subjects in either experiment reported here, rate the importance of all 30 idea units from both texts, using a scale ranging from 1 (*not important at all*) to 5 (*very important*). The average importance rating was calculated for each idea unit, and the intraclass correlation among the average ratings was .78, indicating good interrater reliability (Shrout & Fleiss, 1979). If students selectively included important ideas in the concept map condition, then the average importance rating of recalled ideas should be greater in the concept map condition than in the paragraph recall condition. The results of our analysis confirmed this: Students tended to output ideas with higher normative importance ratings in the concept map condition ( $M = 3.86$ ,  $SE = 0.02$ ) than in the paragraph condition ( $M = 3.76$ ,  $SE = 0.02$ ),  $t(31) = 3.20$ ,  $d = 0.57$ , 95% CI [0.19, 0.94]. Although the raw mean difference was small, the result was robust: for 28 of 32 students (88%), the mean normative importance of recalled ideas was greater in the concept map condition than in the paragraph condition. This analysis indicates that students might have covertly retrieved the same number of ideas in both retrieval practice conditions (which would still benefit learning; see Smith, Roediger, & Karpicke, 2013), but

students chose to include the relatively more important ideas when creating their concept maps.

## Discussion

Experiment 1 showed that practicing retrieval in paragraph format or in concept map format produced approximately equivalent levels of performance on a delayed assessment of learning. Students also gave nearly identical subjective ratings to the two retrieval practice formats (judgments of learning and ratings of enjoyment, difficulty, and interestingness of the tasks), though students did tend to prefer the paragraph retrieval format relative to the concept map format. These results provide preliminary evidence that concept mapping may be an effective retrieval practice activity. Experiment 2 was carried out as a further investigation of the paragraph and concept map formats when used as either retrieval practice or repeated study activities.

## Experiment 2

Experiment 2 was designed with two main purposes in mind. First, we sought to replicate Experiment 1 and generalize the

Table 2

*Students' Ratings of Enjoyment, Difficulty, and Interestingness of the Learning Activities in Experiments 1 and 2*

Learning activity	Experiment 1			Experiment 2		
	Enjoyment	Difficulty	Interest	Enjoyment	Difficulty	Interest
Retrieval practice (no text)						
Concept map	.49 (.04)	.46 (.04)	.49 (.04)	.39 (.06)	.55 (.05)	.42 (.06)
Paragraph	.51 (.04)	.47 (.04)	.53 (.04)	.40 (.04)	.54 (.06)	.49 (.04)
Repeated study (text)						
Concept map	—	—	—	.50 (.06)	.40 (.04)	.55 (.05)
Paragraph	—	—	—	.29 (.06)	.30 (.04)	.32 (.06)

Note. Students' ratings were indicated on a scale from 0 (*not at all*) to 100 (*totally*). Ratings were then converted to proportions. Standard errors of the means are shown in parentheses.



Table 3  
*Joint Probabilities Between Initial Performance and Final Short-Answer Performance in Experiments 1 and 2*

Learning activity	Experiment 1				Experiment 2			
	C <sub>1</sub> C <sub>2</sub>	C <sub>1</sub> N <sub>2</sub>	N <sub>1</sub> C <sub>2</sub>	N <sub>1</sub> N <sub>2</sub>	C <sub>1</sub> C <sub>2</sub>	C <sub>1</sub> N <sub>2</sub>	N <sub>1</sub> C <sub>2</sub>	N <sub>1</sub> N <sub>2</sub>
Retrieval practice (no text)								
Concept map	.34 (.05)	.08 (.02)	.38 (.04)	.20 (.05)	.20 (.04)	.13 (.02)	.38 (.04)	.29 (.05)
Paragraph	.51 (.05)	.11 (.02)	.29 (.02)	.09 (.03)	.28 (.04)	.15 (.01)	.31 (.02)	.26 (.04)
Repeated study (text)								
Concept map	—	—	—	—	.28 (.04)	.26 (.03)	.24 (.02)	.21 (.03)
Paragraph	—	—	—	—	.25 (.04)	.31 (.02)	.18 (.02)	.26 (.04)

*Note.* Standard errors of the means are shown in parentheses. C<sub>1</sub> = items produced during the initial learning activity; N<sub>1</sub> = items that were not produced during the initial learning activity; C<sub>2</sub> = questions correctly answered on the final short answer test; N<sub>2</sub> = questions not correctly answered on the final short answer test.

results to a new set of text materials. Second, we included two new conditions to directly compare concept mapping and paragraph formats when they are used as retrieval practice activities (without the texts) with when they are used as repeated study activities (with the texts present; see Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Agarwal & Roediger, 2011). Thus, in Experiment 2, we factorially crossed the presence of the material during the learning activity (text vs. no text) with the format of the learning activity (concept map vs. paragraph format). Our prediction was that the retrieval-based learning conditions would enhance long-term retention more than the repeated study conditions, even though students in the two conditions completed the exact same activities either with or without the materials in front of them. This result would support the idea that practicing retrieval, rather than the mere act of writing down the material in paragraph or concept map format, is the key to promoting long-term learning.

Method

**Subjects.** Eighty Purdue University undergraduates participated in partial fulfillment of course requirements. None of the students had participated in Experiment 1.

**Materials.** Two science texts were based on information in Stabler, Metz, and Gier (2011). One text, “Enzymes,” had a generalization structure (Meyer, 1975), which means the sentences in the passage provided clarification or examples of one main idea. The other text, “Domains of Life,” had an enumeration structure (like the “Make-Up of Human Blood” text used in Experiment 1). The texts were 283 and 282 words in length, respectively.

**Design.** A 2 (activity format: concept map vs. paragraph) × 2 (learning condition: repeated study vs. retrieval practice) between-subjects design was used. There were four conditions, and 20 students were assigned to each condition. Each student completed the same activity for two texts, and the order in which the texts were presented was held constant across students.

**Procedure.** The procedure was similar to the one used in Experiment 1. Students were tested in small groups in two sessions, and each student was assigned to one of four learning conditions: (a) repeated study–concept map, (b) repeated study–paragraph, (c) retrieval practice–concept map, and (d) retrieval practice–paragraph. During the learning phase, students read one text for 5 min, engaged in a learning activity for 10 min, reread the text for 5 min, and completed the learning activity again for 10 min. Students then repeated the procedure for the second text. All

instructions were identical in the repeated study and retrieval practice conditions, and the total amount of learning time was equivalent in all conditions. The only difference was that in the repeated study conditions, students viewed the texts while they completed the learning activities, whereas in the retrieval practice conditions the students completed the activities without the texts (as in Experiment 1). Thus, students in the repeated study–concept map condition completed their concept maps while reading the texts (Karpicke & Blunt, 2011), and students in the repeated study–paragraph condition were instructed to write everything from the text on their paper in paragraph format (essentially copying the text). In both conditions, students were told to include all of the ideas from the texts. Texts were presented on the computer screen, and students completed the concept mapping or paragraph activities on paper. The subjective rating procedures and the final short answer test procedures were identical to those used in Experiment 1.

Results

An initial analysis indicated that there were no differences among the counterbalancing orders, and the levels of performance and patterns of results were the same for the two texts. Thus, the results have been collapsed across counterbalancing orders and texts.

**Scoring.** The texts were divided into 40 idea units for scoring purposes, and the scoring procedure used in Experiment 1 was used in Experiment 2. Two independent raters scored all recall protocols and short-answer tests, and a third rater resolved all discrepancies to achieve 100% agreement.

**Learning performance.** The right portion of Table 1 shows the mean proportion of idea units produced in each period in the initial learning phase in Experiment 2. Collapsed across conditions, the proportion of ideas produced increased from Period 1 to Period 2 (.38 vs. .52),  $t(79) = 11.59$ ,  $d = 1.33$ , 95% CI [1.02, 1.62]. Students in the repeated study condition (who viewed the texts during the concept map and paragraph activities) produced more ideas than did students in the retrieval practice conditions. This was true for both activity formats in Period 1 (.50 vs. .25),  $t(78) = 8.25$ ,  $d = 1.85$ , 95% CI [1.32, 2.36], and Period 2 (.60 vs. .44),  $t(78) = 5.06$ ,  $d = 1.13$ , 95% CI [0.65, 1.60]. In the repeated study conditions, there were very small differences in the proportion of ideas produced in the concept map and paragraph formats in Period 1 ( $M = 0.48$  vs. 0.53),  $t(38) = 1.07$ ,  $d = 0.34$ , 95% CI

[−0.29, 0.96], or in Period 2 ( $M = 0.58$  vs.  $0.62$ ),  $t(38) = 0.70$ ,  $d = 0.31$ , 95% CI [−0.31, 0.93]. However, as in Experiment 1, students tended to recall more ideas in the paragraph condition than in the concept map condition. There was a small difference in Period 1, (.27 vs. .24),  $t(38) = 0.56$ ,  $d = 0.18$ , 95% CI [−0.44, 0.80], and a larger difference in Period 2, (.48 vs. .39),  $t(38) = 1.96$ ,  $d = 0.62$ , 95% CI [−0.02, 1.25]. In a later section, we report an analysis of the role of idea unit importance in students' performance.

**Final short-answer performance.** Figure 3 shows performance on the final short-answer test 1 week after the initial learning phase. In general, students in the retrieval practice conditions (without the text available) performed better than students in the repeated study conditions (with the text available), but whether the activity was in concept map or paragraph format made little difference for long-term retention.

For verbatim questions, collapsed across activity formats, students in the retrieval practice (no text) conditions outperformed students in the repeated study (with text) conditions (.48 vs. .38),  $t(78) = 2.22$ ,  $d = 0.50$ , 95% CI [0.05, 0.94]. In the retrieval practice condition, there was a small difference between the paragraph and concept map formats, favoring the paragraph format, as was the case in Experiment 1 (.49 vs. .46),  $t(38) = 0.50$ ,  $d = 0.16$ , 95% CI [−0.46, 0.78]. However, in the repeated study condition, there was a larger difference between activity formats, favoring the concept map format over the paragraph format (.43 vs. .33),  $t(38) = 1.59$ ,  $d = 0.50$ , 95% CI [−0.13, 1.13]. This result supports the idea that creating a concept map while studying a text afforded elaborative encoding, as concept mapping enhanced long-term retention relative to essentially copying the text in the repeated study–paragraph condition.

The pattern of results was similar for the inference questions. Collapsed across activity formats, students in the retrieval practice conditions outperformed students in the repeated study conditions (.39 vs. .31),  $t(38) = 2.07$ ,  $d = 0.46$ , 95% CI [0.02, 0.91]. In the retrieval practice condition, there was almost no difference between the paragraph and concept map formats (.40 vs. .39),  $t(38) = 0.27$ ,  $d = 0.09$ , 95% CI [−0.53, 0.70]. Likewise, there was almost no difference between activity formats in the repeated study condition (.30 vs. .32),  $t(38) = 0.30$ ,  $d = 0.09$ , 95% CI [−0.52, 0.71], a result that is somewhat surprising in light of the advantage of concept mapping seen in the verbatim questions, as reported earlier.

**Subjective ratings.** The right panel of Figure 3 shows students' judgments of learning, which were made at the end of each task in the learning phase. Collapsed across activity formats, judgments of learning were higher in the repeated study conditions relative to the retrieval practice conditions (.56 vs. .48),  $t(78) = 1.32$ ,  $d = 0.30$ , 95% CI [−0.15, 0.74]. Although the effect was small in the present experiment, the finding that students believed they had learned more after repeatedly studying than after practicing retrieval is consistent with a wealth of prior work (e.g., Agarwal et al., 2008; Karpicke & Blunt, 2011; see Karpicke, 2012, for review). In the repeated study condition, students' judgments of learning were higher in the concept map condition than in the paragraph condition (.61 vs. .50),  $t(38) = 1.46$ ,  $d = 0.46$ , 95% CI [−0.17, 1.09]. In the retrieval practice condition, the opposite pattern occurred: students' judgments of learning were higher in the paragraph condition than in the concept map condition (.53 vs. .43),  $t(38) = 1.27$ ,  $d = 0.51$ , 95% CI [−0.13, 1.13].

Table 2 shows students' additional ratings of their subjective experiences in the learning tasks, and here we highlight a few findings displayed in the table. Students rated the repeated study–concept map condition as most enjoyable and the repeated study–paragraph task as least enjoyable (.50 vs. .29),  $t(38) = 2.47$ ,  $d = 0.78$ , 95% CI [0.13, 1.42], which is likely due to boredom associated with simply copying the text in the latter condition. The enjoyment ratings of the two retrieval practice conditions fell in between the ratings of the two repeated study conditions. A similar pattern was observed in the interest ratings: students rated the repeated study–concept map task as most interesting and the repeated study–paragraph task as least interesting (.55 vs. .32),  $t(38) = 2.99$ ,  $d = 0.94$ , 95% CI [0.28, 1.59], and the interest ratings of the two retrieval practice conditions fell in between the ratings of the two repeated study conditions. Finally, collapsed across activity formats, the retrieval practice tasks were rated as more difficult than the repeated study tasks (.54 vs. .35),  $t(78) = 3.93$ ,  $d = 0.89$ , 95% CI [0.42, 1.34].

**Conditional analysis.** The right portion of Table 3 shows the results of an analysis of the relationship between initial learning performance and final short-answer performance in Experiment 2. As in Experiment 1, short-answer questions were coded based on the idea unit or units required to answer the questions. Verbatim questions typically required access to a single idea unit ( $M = 1.5$  idea units per verbatim question). For example, the question “What

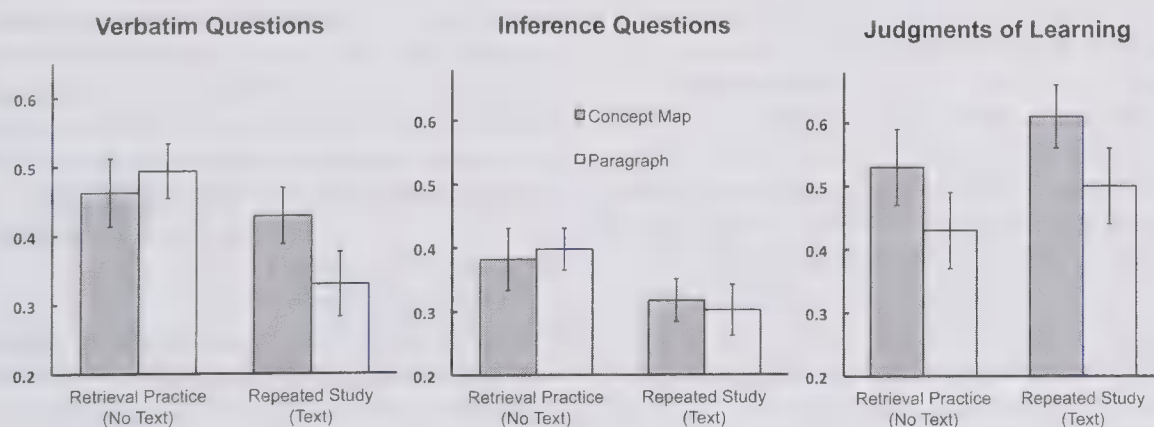


Figure 3. Final short-answer performance for verbatim questions and inference questions (left and middle panels), and judgments of learning (right panel) in Experiment 2. Error bars represent standard errors of the means.



do proteins lose at high temperatures?" corresponded to the idea unit "Proteins lose their structure at high temperatures." Inference questions required access to multiple idea units ( $M = 2.9$  idea units per inference question). For example, the question "What happens to catalytic activity if temperature decreases?" relies on the following idea units: (a) "Catalytic activity is greatly affected by temperature"; (b) "Increasing temperature will also increase the amount of free energy"; (c) "This results in an increased rate of collision"; and (d) "[This] leads to a faster reaction time."

First, we analyzed the relationship between initial learning performance and final short-answer performance, collapsing across activity format (concept map vs. paragraph). As shown in Table 3, the probability of recalling an idea but then failing to answer a short-answer question (intertest forgetting;  $C_1N_2$ ) was greater in restudy conditions than in the retrieval practice conditions,  $t(78) = 7.25$ ,  $d = 1.62$ , 95% CI [1.11, 2.12]. Likewise, the probability of not recalling an idea but then correctly answering a final question ( $N_1C_2$ ) was greater in retrieval practice conditions than in restudy conditions,  $t(78) = 4.99$ ,  $d = 1.12$ , 95% CI [0.64, 1.58]. There were small differences across conditions in  $C_1C_2$ ,  $t(78) = 0.60$ ,  $d = 0.13$ , 95% CI [-0.31, 0.57], and  $N_1N_2$ ,  $t(78) = 0.88$ ,  $d = 0.20$ , 95% CI [-0.24, 0.64].

The pattern of results within the retrieval practice conditions (comparing the concept map format with the paragraph format) replicated the results of Experiment 1. The joint probability of recalling an idea initially and correctly answering a final short-answer question ( $C_1C_2$ ) was slightly greater in the paragraph condition than in the concept map condition,  $t(38) = 1.46$ ,  $d = 0.46$ , 95% CI [-0.17, 1.09]. Likewise, the probability of not recalling an idea but then correctly answering a final question ( $N_1C_2$ ) was slightly greater in the concept map condition than in the paragraph condition,  $t(38) = 1.63$ ,  $d = 0.52$ , 95% CI [-0.12, 1.14]. There was a small difference in intertest forgetting (the probability of recalling an idea but then failing to answer a short-answer question;  $C_1N_2$ ) across conditions, with those in the paragraph condition showing slightly less forgetting,  $t(38) = 0.91$ ,  $d = 0.29$ , 95% CI [-0.34, 0.91]. Finally, the proportion of ideas not recalled or expressed on either test ( $N_1N_2$ ) was slightly greater in the concept map condition relative to the paragraph condition,  $t(38) = 0.93$ ,  $d = 0.29$ , 95% CI [-0.33, 0.92].

**Initial recall and normative importance.** As in Experiment 1, 16 independent raters, who had not served as raters or subjects in Experiments 1 or 2, rated the importance of each idea unit in the two texts used in Experiment 2, using a scale from 1 (*not important at all*) to 5 (*very important*). The average importance rating was calculated for each idea unit, and the intraclass correlation among the average ratings was .80. In the retrieval practice condition, the mean importance rating of the idea units that students recalled was greater in the concept map condition ( $M = 3.60$ ,  $SE = 0.02$ ) than in the paragraph condition ( $M = 3.48$ ,  $SE = 0.03$ ),  $t(38) = 3.00$ ,  $d = 0.95$ , 95% CI [0.29, 1.60]. However, in the repeated study condition, there was a smaller difference between the mean importance ratings in the concept map ( $M = 3.54$ ,  $SE = 0.02$ ) and paragraph conditions ( $M = 3.50$ ,  $SE = 0.02$ ),  $t(38) = 1.36$ ,  $d = 0.43$ , 95% CI [-0.20, 1.05]. Thus, as in Experiment 1, when students practiced retrieval, they tended to include ideas with higher normative importance ratings in the concept map conditions than in the paragraph conditions, though this difference was much

smaller when students completed the activities with the materials in front of them.

## Discussion

Experiment 2 showed that actively retrieving material during learning, either by creating concept maps or by writing the material in paragraph format, enhanced long-term retention more than completing the same activities in the presence of the materials (as study activities). Practicing retrieval produced more learning than repeated studying even though students re-experienced the entire set of material in the repeated study conditions, whereas students only re-experienced what they could recall in the retrieval practice conditions. Indeed, the proportion of ideas recalled in the retrieval practice conditions was lower than the proportion of ideas produced on the concept map or paragraph protocols in the repeated study conditions. It is important to note that the concept map and paragraph formats were equally effective as retrieval practice activities. As in Experiment 1, there were no additional benefits conferred by retrieval-based concept mapping beyond practicing retrieval in paragraph format. There was a small cost to retrieval-based concept mapping in the initial recall periods, on which students recalled fewer ideas in the concept map condition than in the paragraph condition. However, this cost was not seen on the final delayed assessments of long-term retention. Together with Experiment 1, the results of Experiment 2 show that concept mapping can serve as an effective learning task when it is implemented as a retrieval-based learning activity.

## General Discussion

The purpose of the present experiments was to examine the effectiveness of retrieval-based concept mapping. The results show that the critical factor in retrieval-based learning is requiring students to think back to and recall material, while the format in which information is retrieved (concept map or paragraph format) did not much matter. We review three important findings from the present experiments in light of hypotheses proposed in the introduction.

First, concept mapping and paragraph formats were equally effective retrieval-based learning activities. When students created retrieval-based concept maps of the materials, there were no practical differences, relative to recalling in paragraph format, on delayed short-answer performance in Experiment 1 or 2. Furthermore, Experiment 2 showed that both activity formats produced retrieval practice effects: Students performed better on a final test when the initial activities required retrieval (in the absence of the texts) rather than studying or elaborating on the material (in the presence of the texts). This advantage of retrieval practice occurred even though students in the retrieval conditions produced less material during the initial learning activities relative to students in the repeated study conditions.

Second, retrieving in paragraph format produced greater long-term performance relative to restudying and rewriting the material in paragraph format. It is reasonable to wonder whether the locus of retrieval practice effects rests in the act of writing itself, rather than in the mental activity of retrieving and reconstructing knowledge. If this were the case, the repeated study-paragraph condition in Experiment 2 should have produced long-term performance

similar to that produced by the retrieval practice–paragraph condition. Indeed, because students were able to re-experience the entire set of material in the repeated study condition, one might expect that condition to outperform the retrieval practice condition. However, the opposite result occurred in Experiment 2, confirming that the act of retrieving knowledge itself, rather than the act of writing, drives the benefits seen in retrieval-based learning activities.

Third, students generally believed they had learned more after repeatedly studying than after practicing retrieval. This result is consistent with a wealth of prior research (see Karpicke, 2012) and is also broadly consistent with a cue utilization approach to metacognitive judgments (e.g., Koriat, 1997). According to this view, students base their judgments of learning in part on the ease of processing they experience during a learning activity. When students complete activities with the text in front of them, processing is fluent and easy, whereas when students complete activities without the text, they base their judgments on the ease or difficulty with which the material can be brought back to mind during retrieval. Thus, repeated study activities tend to afford overconfident judgments of learning, whereas retrieval practice leads to underconfident judgments. In Experiment 2, students rated concept mapping as more interesting and enjoyable than studying by copying the text in paragraph form, but students' ratings did not differ among concept map and paragraph formats when completed as retrieval activities. Despite some speculation that concept mapping might somehow promote or improve metacognitive performance (e.g., Mintzes et al., 1997), the present experiments offer no evidence that this is true (see too Karpicke & Blunt, 2011).

The key finding from the present experiments was that retrieval practice was equally effective when done in concept map or paragraph format. Students did not gain additional benefits by retrieving knowledge in concept map format relative to retrieving in paragraph format. Concept mapping is assumed to promote organizational or relational processing that should improve learning, but our results are consistent with the possibility that such organizational processing may be redundant with the processing people already engage in when practicing retrieval in other ways. Furthermore, practicing retrieval in concept map format did not impair learning relative to practicing retrieval in paragraph format. This finding suggests that the concept mapping task did not introduce extra cognitive load or divide attention in ways that were detrimental to learning. When students retrieved in concept map format, they tended to recall fewer ideas than when they retrieved in paragraph format, because they selectively reported ideas that were rated as most important. However, this was not detrimental to long-term learning either. Thus, the present experiments support the conclusion that concept mapping can indeed function as an effective learning activity when it involves practicing retrieval.

## Conclusion

Retrieval practice is a powerful way to enhance long-term meaningful learning of educationally relevant content. The present results show that practicing retrieval, either by creating concept maps or by writing down the material in paragraph format, enhanced long-term learning more than completing the same tasks as study activities. The locus of these learning effects was in the act of retrieving knowledge, rather than the mere act of writing down

the material in paragraph or concept map format. It is important to note that the results show that concept mapping can indeed serve as an effective task when it is implemented as a retrieval-based learning activity. The key element for promoting meaningful learning was not the format of the activity; it was the requirement to engage in active retrieval practice during learning.

## References

- Agarwal, P. K., Karpicke, J. D., Kang, S. H., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, 22, 861–876. doi:10.1002/acp.1391
- Agarwal, P. K., & Roediger, H. L., III. (2011). Expectancy of an open-book test decreases performance on a delayed closed-book test. *Memory*, 19, 836–852.
- Cook, L. K., & Mayer, R. E. (1988). Teaching readers about the structure of scientific text. *Journal of Educational Psychology*, 80, 448–456.
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.
- Hay, D., Kinchin, I., & Lygo-Baker, S. (2008). Making learning visible: The role of concept mapping in higher education. *Studies in Higher Education*, 33, 295–311. doi:10.1080/03075070802049251
- Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, 21, 157–163. doi:10.1177/0963721412443552
- Karpicke, J. D., & Blunt, J. R. (2011, February 11). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331, 772–775. doi:10.1126/science.1199327
- Karpicke, J. D., Lehman, M., & Aue, W. R. (in press). Retrieval-based learning: An episodic context account. In B. Ross (Ed.), *The psychology of learning and motivation*. New York, NY: Elsevier.
- Karpicke, J. D., & Roediger, H. L., III. (2010). Is expanding retrieval a superior method for learning text materials? *Memory & Cognition*, 38, 116–124. doi:10.3758/MC.38.1.116
- Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language*, 67, 17–29. doi:10.1016/j.jml.2012.02.004
- Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language*, 62, 227–239. doi:10.1016/j.jml.2009.11.010
- Kelley, K. (2007). Methods for the behavioral, educational, and social science: An R package. *Behavior Research Methods*, 39, 979–984. doi:10.3758/BF03192993
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126, 349–370. doi:10.1037/0096-3445.126.4.349
- Meyer, B. J. (1975). *The organization of prose and its effects on memory*. Amsterdam, the Netherlands: North-Holland.
- Mintzes, J. J., Wandersee, J. H., & Novak, J. D. (1997). Meaningful learning in science: The human constructivist perspective. In G. D. Phye (Ed.), *Handbook of academic learning* (pp. 405–447). Orlando, FL: Academic Press. doi:10.1016/B978-012554255-5/50014-4
- Novak, J. D. (1976). Understanding the learning process and effectiveness of teaching methods in the classroom, laboratory, and field. *Science Education*, 60, 493–512. doi:10.1002/sce.3730600410
- Novak, J. D. (2005). Results and implications of a 12-year longitudinal study of science concept learning. *Research in Science Education*, 35, 23–40. doi:10.1007/s11165-004-3431-4
- Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn*. New York, NY: Cambridge University Press. doi:10.1017/CBO9781139173469



- Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17, 249–255. doi:10.1111/j.1467-9280.2006.01693.x
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.
- Smith, M. A., Roediger, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39, 1712–1725.
- Stabler, L. B., Metz, M., & Gier, P. (2011). *AP Biology 2012*. New York, NY: Kaplan.
- Stewart, J., Van Kirk, J., & Rowell, R. (1979). Concept maps: A tool for use in biology teaching. *The American Biology Teacher*, 41, 171–175. doi:10.2307/4446530
- Tulving, E. (1964). Intratrial and intertrial retention: Notes toward a theory of free recall verbal learning. *Psychological Review*, 71, 219–237. doi:10.1037/h0043186
- Tulving, E. (1983). *Elements of episodic memory*. New York, NY: Oxford University Press.
- Vanides, J., Yin, Y., Tomita, M., & Ruiz-Primo, M. A. (2005). Using concept maps in the science classroom. *Science Scope*, 28, 27–31.

## Appendix

### Examples of Verbatim and Inference Questions Used in Experiment 1<sup>a</sup> and Experiment 2

#### Experiment 1. Sample questions from text on “Make-Up of Human Blood”:

Verbatim question:

“What happens when hemoglobin combines with oxygen?”

(Sample answer: Oxygen is released to cells in the body.)

Inference question:

“What would happen to blood flow from a wound if the body did not have fibrin?”

(Sample answer: Blood would not clot, because fibrin is needed to form a meshwork of fibers that trap blood cells and aid in clotting.)

#### Experiment 2. Sample questions from text on “Enzymes”:

Verbatim question:

“What are two forms of free energy?”

(Sample answer: Heat and kinetic energy.)

Inference question:

“What happens to catalytic activity if temperature decreases?”

(Sample answer: Catalytic activity decreases because increasing temperature increases the rate of molecular collision, which leads to a faster reaction time.)

<sup>a</sup> For a complete set of questions, see Karpicke & Blunt (2011).

Received June 7, 2013

Revision received October 21, 2013

Accepted December 29, 2013 ■

# Can Parents' Involvement in Children's Education Offset the Effects of Early Insensitivity on Academic Functioning?

Jennifer D. Monti and Eva M. Pomerantz  
University of Illinois at Urbana–Champaign

Glenn I. Roisman  
University of Minnesota

Data from the National Institute of Child Health and Human Development Study of Early Child Care and Youth Development ( $N = 1,312$ ) were analyzed to examine whether the adverse effects of early insensitive parenting on children's academic functioning can be offset by parents' later involvement in children's education. Observations of mothers' early insensitivity (i.e., 6–54 months) interacted with teachers' reports of parents' later involvement (i.e., 1st–5th grade) in predicting children's academic functioning as reflected in observed classroom engagement and performance on standardized achievement tests at the end of elementary school (i.e., 5th grade): Although mothers' insensitivity foreshadowed dampened academic functioning among children when parents' involvement was relatively low, it did not do so when parents' involvement was average or higher.

**Keywords:** academics, achievement, parent involvement, parent sensitivity, parenting

Insensitive parenting (i.e., unresponsiveness, hostility, and intrusiveness) early in children's lives appears to undermine children's engagement and achievement in the academic arena not only when children first enter school but also throughout childhood, adolescence, and even into adulthood (e.g., Fraley, Roisman, & Haltigan, 2013; Raby, Roisman, Fraley, & Simpson, 2013; Stams, Juffer, & van IJzendoorn, 2002). A key question is whether aspects of children's later environment can offset these costs of early insensitivity. There is much evidence suggesting that parents' involvement in children's education (e.g., volunteering at school, attending parent–teacher conferences, and discussing school with children) is instrumental in promoting children's academic functioning (for a review, see Pomerantz, Moorman, & Cheung, 2012). Hence, such involvement may be an important aspect of children's later environment that can offset the academic problems associated with early insensitive parenting. The goal of the current research was to evaluate whether parents' involvement in children's education can compensate for the adverse effects of early insensitive parenting on children's academic functioning.

## Parents' Early Insensitivity

Insensitive parenting has been conceptualized by the National Institute of Child Health and Human Development [NICHD] Early Child Care Research Network ([ECCRN] 1997, 2004, 2008) and others (e.g., Campbell, Matestic, von Stauffenberg, Mohan, & Kirchner, 2007; Stams et al., 2002) as parents' unresponsiveness to children's nondistress signals, hostility (vs. warmth) toward children, and intrusiveness (vs. autonomy support). Parents' insensitivity has been argued to undermine children's academic functioning through several mechanisms. For example, when parents fail to respond contingently to children and are intrusive, children come to feel helpless in affecting their environment (e.g., Nolen-Hoeksema, Wolfson, Mumme, & Guskin, 1995; Riksen-Walraven, 1978). Thus, children disengage when confronted with challenge in that they are less attentive, self-directed, and persistent, which may interfere with their learning (e.g., Bornstein & Tamis-LeMonda, 1997; Frodi, Bridges, & Grolnick, 1985; NICHD ECCRN, 2008). The case has also been made that when parents are insensitive, children fail to develop a secure attachment (De Wolff & van IJzendoorn, 1997); as a consequence, they do not view their caregiver as a reliable source of support, which undermines their willingness to engage in potentially distressing but cognitively stimulating behaviors such as active exploration of the environment and persistence in the face of challenge (e.g., Bretherton, 1985; Main, 1983; Matas, Arend, & Sroufe, 1978; for additional mechanisms by which attachment contributes to children's academic adjustment, see van IJzendoorn, Dijkstra, & Bus, 1995).

Consistent with these ideas, exposure to early insensitivity appears to disrupt the development of a foundation for children's later academic functioning. For example, in the first year of life, children whose mothers are insensitive in that they are unresponsive or intrusive exhibit dampened attentiveness and persistence (e.g., Bornstein & Tamis-LeMonda, 1997; Frodi, Bridges, & Grolnick, 1985). Moreover, early insensitive parenting predicts poorer cognitive skills during the preschool years, even after taking into

---

This article was published Online First February 17, 2014.

Jennifer D. Monti and Eva M. Pomerantz, Department of Psychology, University of Illinois at Urbana–Champaign; Glenn I. Roisman, Institute of Child Development, University of Minnesota.

We are grateful to the principal investigators, site coordinators, and participants of the National Institute of Child Health and Human Development Study of Early Child Care and Youth Development. We appreciate the constructive comments provided by the members of the Center for Parent-Child Studies at the University of Illinois at Urbana–Champaign on an earlier version of this article.

Correspondence concerning this article should be addressed to Jennifer D. Monti or Eva M. Pomerantz, Department of Psychology, University of Illinois at Urbana–Champaign, 603 East Daniel Street, Champaign, IL 61820. E-mail: schmid41@illinois.edu or pomerantz@illinois.edu



account children's earlier cognitive skills (Lemelin, Tarabulsy, & Provost, 2006). Recent evidence suggests that such effects are not simply due to shared genetics (Roisman & Fraley, 2012). Moreover, the deleterious effects of early insensitive parenting on children's academic functioning appear to endure into adolescence. Using the NICHD Study of Early Child Care and Youth Development (SECCYD), Fraley and colleagues (2013) demonstrated that the predictive significance of mothers' early (i.e., 6–36 months) insensitivity on children's achievement remained relatively constant throughout childhood and into adolescence, even after accounting for the stability of mothers' insensitivity across these years of development as well as potential confounds such as mothers' educational attainment, children's race, and family income. Similar findings were recently observed in the Minnesota Longitudinal Study of Risk and Adaptation through age 32 with a focus on academic attainment (Raby et al., 2013).

### **The Compensatory Role of Parents' Involvement in Children's Education**

By becoming involved in children's education, parents can offset the adverse effects of early insensitive parenting on children's academic functioning. Parents' involvement in children's education has been argued to foster the psychological resources necessary for children's optimal academic functioning (for reviews, see Pomerantz & Moorman, 2010; Pomerantz, Moorman, & Cheung, 2012). For example, such involvement may highlight the value of learning to children, which may heighten their engagement in school, thereby enhancing their achievement (e.g., Epstein, 1988; Grolnick & Slowiaczek, 1994). The case has also been made that by providing additional instruction or opportunities for practice, parents' involvement develops important academic competencies among children (e.g., Cheung & Pomerantz, 2011; Senchal & LeFevre, 2002). Parents' involvement on the school front may lead to enhanced learning among children by increasing the attention children receive from teachers (Epstein & Becker, 1982; for additional mechanisms by which involvement contributes to children's academic adjustment, see Pomerantz et al., 2012). The benefits of parents' involvement are evident even when parents are insensitive as reflected in intrusiveness (Cheung & Pomerantz, 2011).

Because children exposed to insensitive parenting early in their lives are academically at risk—often lacking critical competencies for achievement in this context—they may be in much need of the resources provided by parents' involvement in children's education. Thus, they may be particularly likely to benefit from parents' involvement, such that over time parents' involvement can compensate for the costs of early insensitivity. Suggestive of this possibility, parents' involvement can offset another aspect of children's environment that appears to undermine children's academic functioning. Dearing and colleagues (2006) found that achievement disparities between children of less versus highly educated mothers were moderated by parents' involvement in children's education (see also Dearing, McCartney, Weiss, Kreider, & Simpkins, 2004). Specifically, during the early elementary school years, when parents' school-based involvement was low, children with less educated mothers had poorer literacy achievement than did children with more educated mothers, but when such involvement was high, this difference was not evident. Parents' involvement

may play a similar compensatory role when it comes to early insensitive parenting: When parents are involved in children's education, the academic problems associated with early insensitive parenting may be reduced.

Although many insensitive parents are not involved in children's education, a sizable number are (for associations between insensitivity—as manifest in intrusiveness and hostility—and involvement, see Cheung & Pomerantz, 2011; Pomerantz, Wang, & Ng, 2005; Steinberg, Lamborn, Dornbusch, & Darling, 1992). Sensitive parenting may require a somewhat different set of skills and values than does involvement in children's education. For example, parents' sensitivity entails the capacity to show warmth as well as a concern with children's psychological needs, whereas parents' involvement entails the capacity to monitor children's progress in school as well as a concern with children's performance. Once children enter the formal school system, teachers may elicit parents' involvement on the school front (e.g., communicating with teachers or volunteering at school) via invitations (e.g., Green, Walker, Hoover-Dempsey, & Sandler, 2007). The benefits of school-based involvement may not be appreciably dampened by insensitivity because such involvement requires relatively little interaction between children and parents, but can still provide important resources—for example, by conveying that school is valuable. Indeed, the positive effects of school-based involvement are more consistent than those of home-based involvement (e.g., assisting children with homework and discussing school with children), which almost always entails interaction between children and parents (Pomerantz, Moorman, & Litwack, 2007). Nonetheless, because home-based involvement also provides important resources (e.g., instruction and practice), both home- and school-based involvement may play a compensatory role.

### **Overview of the Current Research**

The key hypothesis guiding this research was that the adverse effects of early insensitive parenting on children's later academic functioning can be offset by parents' involvement in children's education. This notion reflects one of the central tenets of Bronfenbrenner's (1992) ecological systems theory: The influence of children's proximal environment—that is, the microsystem as manifest in parents' early sensitivity—is shaped in part by the broader environment—in this case, the mesosystem as manifest in parents' involvement in children's education given that it includes interactions with school personnel. We focused on parents' involvement during the elementary school years (i.e., first to fifth grade) because there is often substantial opportunity for parents to become involved on the school front at this time. Parents' involvement was assessed with teacher reports explicitly referencing parents' involvement on the school front given that such involvement is not only particularly reflective of the mesosystem but also may be minimally influenced by parental insensitivity (Pomerantz et al., 2007). The measure also asked about parents' involvement in children's education in general, which may capture parents' involvement on the home, as well as school, front.

We analyzed data from the NICHD SECCYD. Observations of mothers' insensitive parenting prior to children's entry into the formal education system (i.e., 6–54 months) were used as measures of early insensitivity. Once children entered first grade,

teachers reported every year on parents' involvement in children's education. The annual reporting allowed us to examine parents' involvement over an extended period of time (i.e., 5 years), which is of import given that years of insensitive parenting prior to children's entry into school are unlikely to be immediately undone by a short phase (e.g., 1 year during first grade) of involvement. The effects of these two aspects of children's environment on children's academic functioning at the end of elementary school (i.e., fifth grade) were examined, adjusting for such functioning as children entered elementary school (i.e., first grade). This allowed us to rule out the possibility that the effects of parents' involvement were due simply to developmental processes in place before children entered elementary school. To identify the breadth of the compensatory role of parents' involvement, we investigated multiple forms of children's academic functioning, which were assessed using diverse methods (i.e., observations of engagement in the classroom, standardized achievement test performance, and teachers' reports of academic competencies).

Two major steps were taken to rule out alternative explanations. First, it is possible that early insensitive parenting followed by parents' involvement in children's education represents a change in insensitivity among parents such that by the elementary school years, parents have become less insensitive. Thus, we adjusted for mothers' insensitivity during elementary school. We also evaluated whether it was insensitivity at this time rather than involvement that offsets the adverse effects of early insensitivity. Second, because parents' involvement in children's education can compensate for low educational attainment among mothers (e.g., Dearing et al., 2006), and insensitive parenting is particularly common when socioeconomic resources are low (e.g., Linver, Brooks-Gunn, & Kohen, 2002; NICHD ECCRN, 2005), we examined whether the interactive effects of early insensitivity and elementary school involvement are unique or simply reflect the interactive effects of educational attainment and involvement.

## Method

### Participants

Families were recruited for the NICHD SECCYD from hospitals in 10 locations (Little Rock, AR; Irvine, CA; Lawrence, KS; Boston, MA; Philadelphia, PA; Pittsburgh, PA; Charlottesville, VA; Morganton, NC; Seattle, WA; Madison, WI) shortly after mothers gave birth (for sampling and recruitment details, see <http://secc.rti.org>). The resulting sample consisted of 1,364 children and their mothers. The current analyses used assessments from Phases 1 (birth to 3 years), 2 (54 months to first grade), and 3 (second to sixth grade). The analytic sample was restricted to dyads with data available for at least one of the assessments of maternal insensitivity or parental involvement examined in the current analyses ( $N = 1,312$ ). This sample was predominantly (81%) White (13% African American, 2% Asian, and 5% other minority groups); 52% of children were boys. At Phase 1, there was a range of educational attainment among mothers: 30% had a high school degree or less, 55% had completed some college or earned a college degree, and 15% had completed some graduate work or earned a graduate degree. Most (84%) families had income-to-needs ratios classified as not poor (i.e.,  $\geq 1$ ).

## Measures

Table 1 presents descriptive statistics for the measures used in the current report.

**Maternal insensitivity.** *Early maternal insensitivity* was assessed in the context of mother-child interactions during semi-structured play. Mothers were provided age-appropriate toys designed to elicit joint play and instructed to play with children using the toys in a specific order, or to use the toys to accomplish a specific task (e.g., completing a maze on an etch-a-sketch toy). The interactions were videotaped in the home at 6 and 15 months and in the laboratory at 24, 36, and 54 months. At the 6-, 15-, and 24-month observations, mothers' behavior was coded for sensitivity to nondistress (i.e., responsiveness to child's signals), positive regard (i.e., expressions conveying positive feelings toward child), and intrusiveness (i.e., controlling behaviors) using a 4-point scale (from 1 = *not at all characteristic of the interaction* to 4 = *highly characteristic of the interaction*). Maternal insensitivity was computed as the sum of the three ratings, with sensitivity to nondistress and positive regard reverse scored ( $\alpha s = .70-.79$ ). At the 36- and 54-month observations, mothers' behavior was coded for supportive presence (i.e., positive assistance and emotional support), hostility (i.e., angry or rejecting behaviors), and respect for auton-

Table 1  
*Descriptive Statistics*

Measure	<i>M (SD)</i>	Observed range
Early maternal insensitivity		
6 months	9.21 (1.78)	3-12
15 months	9.40 (1.65)	3-12
24 months	9.35 (1.76)	3-12
36 months	17.19 (2.78)	4-21
54 months	16.95 (2.91)	4-21
Elementary school maternal insensitivity		
1st grade	16.88 (3.03)	5-21
3rd grade	16.34 (2.49)	4-21
5th grade	16.50 (2.42)	7-21
Elementary school parent involvement		
1st grade	3.55 (.98)	1-5
2nd grade	3.55 (.96)	1-5
3rd grade	3.33 (.94)	1-5
4th grade	3.30 (.92)	1-5
5th grade	3.25 (.94)	1-5
Elementary school classroom engagement		
1st grade	55.91 (4.73)	28-60
5th grade	40.92 (8.43)	11.25-59.00
Elementary school Woodcock-Johnson achievement tests		
1st grade	477.15 (10.82)	432.00-509.71
5th grade	508.46 (11.66)	417.50-540.60
Elementary school academic skills		
1st grade	3.28 (.90)	1-5
5th grade	3.46 (.85)	1-5
Elementary school performance		
1st grade current	3.41 (.84)	1-5
5th grade current	3.48 (.96)	1-5
Covariates		
Maternal education	14.28 (2.50)	7-21
Income-to-needs ratio	3.60 (2.85)	0.15-27.36
Maternal depression	9.36 (6.76)	0-43



omy (i.e., acknowledgement and support of child's intentions and independence) using a 7-point scale (from 1 = *not at all characteristic of the interaction* to 7 = *highly characteristic of the interaction*). Maternal insensitivity was computed as the sum of the three scales, with supportive presence and respect for autonomy reverse scored ( $\alpha$ s = .78–.84 interrater reliability:  $r$ s = .71–.87,  $ps < .001$ ). The associations among maternal insensitivity across the five measurement points were sizable ( $r$ s = .30–.52,  $ps < .001$ ); together, the five formed a reliable composite score ( $\alpha$  = .76). Thus, a single index was created by taking the mean of mothers' standardized insensitivity scores prior to children's formal schooling (i.e., at 6, 15, 24, 36, and 54 months), with higher numbers reflecting heightened maternal insensitivity.

*Elementary school maternal insensitivity* was assessed when children were in first, third, and fifth grade during videotaped sessions in the laboratory. In first grade, insensitivity was assessed in the context of a semistructured mother–child play task. In third and fifth grades, insensitivity was assessed in the context of mothers and children discussing areas of disagreement and working on a planning task. Mothers' behavior was coded for supportive presence, hostility, and respect for autonomy (interrater reliability:  $r$ s = .72–.83). At each time point, the sum of the three scales was taken, with supportive presence and respect for autonomy reverse scored ( $\alpha$  = .80–.85). A single index was created by taking the mean of mothers' standardized insensitivity scores at the first, third, and fifth grades, which were sizably associated with one another ( $r$ s = .43–.47,  $ps < .001$ ;  $\alpha$  = .71).

**Parental involvement.** Teachers completed the Parent-Teacher Involvement Questionnaire (Kohl, Lengua, McMahon, & Conduct Problems Prevention Research Group, 2000; Miller-Johnson, Maumary-Gremaud, & Conduct Disorders Research Group, 1995) each year from the time children were in first through fifth grade. Four of the 10 items comprising the questionnaire ask about parents' involvement in children's education explicitly on the school front (e.g., "How often does this parent volunteer or visit at school?" and "How often does this parent send things to class like story books or objects?"); four ask about parents' involvement more generally (e.g., "How involved is this parent in his/her child's education and school life?" and "How important is education in this family?"). Two items do not directly assess parental involvement, but rather the relationship between parents and teachers ("How well do you feel you can talk to and be heard by this parent?" and "If you had a problem with this child, how comfortable would you feel talking to his/her parent about it?"); thus, they were omitted for the current analyses. Teachers rated items on a 5-point scale (from 1 = *not at all* to 5 = *a great deal*). Scores were computed by taking the mean of the eight items assessing parental involvement ( $\alpha$ s = .91–.92), with higher scores indicating heightened involvement. A single index was created by taking the mean of the scores from first through fifth grades, which were sizably associated ( $r$ s = .54–.64,  $ps < .001$ ) and together formed a reliable composite ( $\alpha$  = .88).

**Child academic functioning.** Multiple forms of children's academic functioning were examined. We used assessments at the beginning (i.e., first grade) and end (i.e., fifth grade) of elementary school. *Observations of classroom engagement* were made using the Classroom Observation System, a rating system created specifically for the NICHD SECCYD. Observers recorded the frequency of children's behaviors in 10-min periods consisting of

30-s observe, 30-s record intervals. In first grade, children's engagement was computed as the sum of the active and passive engagement scales (interrater reliability:  $r$  = .82,  $p < .001$ ), which were based on six observe-and-record 10-min periods; in fifth grade, children's engagement was computed as the sum of the engaged in learning and highly engaged scales, which were based on eight to 11 observe-and-record 10-min periods (interrater reliability:  $r$  = .97,  $p < .001$ ). Higher scores indicate heightened classroom engagement.

Children's achievement was assessed with the *Woodcock-Johnson Tests of Psychoeducational Achievement-Revised* (Woodcock & Johnson, 1989). In first grade, children completed four subtests of cognitive aptitude: Memory for Names (long-term retrieval), Memory for Sentences (short-term memory), Incomplete Words (auditory processing), and Picture Vocabulary (verbal comprehension). Children also completed three subtests of achievement: Letter-Word Identification (learning and reading), Applied Problems (mathematical and practical problem solving), and Word Attack (phonic and structural analysis). In fifth grade, children completed one subtest of cognitive ability (i.e., Picture Vocabulary) and three subtests of achievement (i.e., Letter-Word Identification, Applied Problems, and Passage Comprehension). Raw scores for each test were converted to W scores, a transformation of the Rasch ability scale centered at the value of 500. The mean of the standardized W scores on the subtests was taken at the first ( $r$ s = .25–.85,  $ps < .001$ ) and fifth ( $r$ s = .44–.74,  $ps < .001$ ) grades, with higher numbers indicating higher achievement.

*Teacher reports of academic competencies* when children were in the first and fifth grades were used. Teachers completed the Academic Skills questionnaire from the Early Childhood Longitudinal Study (Nicholson, Atkins-Burnett, & Meisels, n.d.). In first grade, teachers rated children's ability to perform 25 age-appropriate skills on a 5-point scale (from 1 = *not yet* to 5 = *proficient*;  $\alpha$  = .97). Fifteen items captured language and literacy skills (e.g., "Reads first grade books fluently"), and 10 captured mathematical thinking skills (e.g., "Understands place values"). In fifth grade, teachers reported on a comparable set of 23 skills ( $\alpha$  = .95). Ten items captured language and literacy skills (e.g., "Conveys ideas clearly") and 13 captured mathematical thinking skills (e.g., "Uses a variety of strategies solving math problems"). Scores were computed as the mean of the items, with higher scores indicating higher proficiency. Teachers also completed the Current School Performance subscale of the Child Evaluation questionnaire (Pierce, Hamm, & Vandell, 1999) when children were in the first and fifth grades. Teachers rated children's performance in reading, oral language, written language, math, social studies, and science on a 5-point scale (from 1 = *below grade level* to 5 = *excellent*;  $\alpha$ s = .93 and .95). At both the first and fifth grades, the Academic Skills and Current School Performance scales were substantially associated ( $r$ s = .77 and .78,  $ps < .001$ ); thus, a composite index was created by taking the mean of the two, with higher numbers reflecting greater competence.

**Covariates.** Mothers reported their educational attainment (i.e., years of education), child gender, and child race (i.e., American Indian, Asian, Black, White, or "other") when children were 1 month old. Ongoing assessments of family income-to-needs ratio and maternal depression were completed throughout the study. To compute income-to-needs ratios, family total income was divided by the poverty threshold based on family size. Maternal depression

was measured with the 20-item Center for Epidemiology Depression Scale (Radloff, 1977). To capture the same time frame as the assessments of mothers' early sensitivity, we averaged over the assessments when children were 6, 15, 24, 36, and 54 months old for family income-to-needs ratio and maternal depression ( $\alpha s = .89-.91$ ).

## Results

### Missing Values

Due to failure to complete all assessments and item nonresponse, the NICHD SECCYD contains incomplete data. Analyses were conducted not only on the original data set with listwise deletion but also on 25 imputed data sets created with multiple imputation (Rubin, 1987; Schafer & Graham, 2002) using the fully conditional specification method in SPSS 19.0 (IBM Corp., 2010). The imputation model included all of the variables in the central and supplementary analyses, as well as the interaction terms as suggested by Enders (2010). Results from the imputed data sets were combined, yielding pooled estimates that account for variability in estimates between and within the data sets. The pattern of results was practically identical for the estimates based on the original and imputed data sets (see Table 3). Results from the original data set are reported in the text (see also Tables 1 and 2 and Figures 1 and 2), with the results from both the original and imputed data sets presented in Table 3.

### Preliminary Analyses

As shown in Table 2, mothers' insensitivity—both early in children's lives (i.e., 6–54 months) and during elementary school (i.e., first, third, and fifth grade)—was inversely associated with parents' involvement in children's education during elementary school (i.e., first to fifth grade). The associations were sizable ( $r s = -.49$  and  $-.43$ ,  $p s < .001$ ), but far from unity. Indeed, 17% ( $n = 187$ ) of the sample of 1,109 with relevant data were above the 50th percentile on both early insensitivity and elementary school involvement, 33% ( $n = 369$ ) were above the 50th percentile on early insensitivity and below the 50th percentile on elementary school involvement, 34% ( $n = 377$ ) were below the 50th percentile on early insensitivity and above the 50th percentile on elementary school involvement, and 16% ( $n = 176$ ) were below the 50th percentile on both early insensitivity and elementary school involvement.

Consistent with prior research, mothers' early, as well as elementary school, insensitivity was associated with children's academic functioning at both first and fifth grade (e.g., mothers' early insensitivity was inversely associated with children's standardized achievement test performance at fifth grade). Also consistent with prior research, parents' elementary school involvement was associated with heightened academic functioning among children at both the beginning (i.e., first grade) and end (i.e., fifth grade) of elementary school. The covariates (e.g., maternal education and depression) were generally associated with parenting and children's academic functioning, indicating the importance of taking them into account.

Table 2  
Correlations

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Early insensitivity	—													
2. Elementary school insensitivity	.62***	—												
3. Elementary school involvement	-.49***	-.43***	—											
4. 1st-grade classroom engagement	-.13***	-.15***	.15***	—										
5. 5th-grade classroom engagement	-.22***	-.17***	.21***	.13***	—									
6. 1st-grade Woodcock-Johnson	-.42***	-.38***	.40***	.11*	.20***	—								
7. 5th-grade Woodcock-Johnson	-.46***	-.43***	.43***	.07^	.25***	.81***	—							
8. 1st-grade academic competencies	-.35***	-.30***	.44***	.16***	.20***	.69***	.65***	—						
9. 5th-grade academic competencies	-.43***	-.39***	.51***	.15***	.30***	.66***	.72***	.65***	—					
10. Child gender	-.08**	-.08**	.04	.15***	.12**	-.02	.01	.06^	.12***	—				
11. Child race	.39***	.34***	-.31***	-.06^	-.16***	-.28***	-.32***	-.18***	-.26***	.00	—			
12. Maternal education	-.50***	-.43***	.50***	.11***	.14***	.42***	.45***	.35***	.42***	.03	-.20***	—		
13. Income-to-needs ratio	-.40***	-.35***	.41***	.08**	.20***	.34***	.35***	.24***	.33***	.04	-.23***	.55***	—	
14. Maternal depression	.32***	.30***	-.31***	-.06*	-.14***	-.25***	-.24***	-.21***	-.23***	.02	.19***	-.31***	-.29***	—

Note. For child gender, 1 = male and 2 = female; for child race, 1 = White and 2 = non-White. The estimates are from the original data set.  
^  $p < .10$ . \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .



Table 3

*Predicting Children's Fifth-Grade Academic Functioning From Mothers' Early Sensitivity and Parents' Elementary School Involvement*

Predictor variable	Observations of classroom engagement				Woodcock-Johnson achievement tests				Teacher reports of academic competence			
	Original data		Multiple imputation		Original data		Multiple imputation		Original data		Multiple imputation	
	$\beta$	$\Delta R^2$	<i>B</i>	<i>SE</i>	$\beta$	$\Delta R^2$	<i>B</i>	<i>SE</i>	$\beta$	$\Delta R^2$	<i>B</i>	<i>SE</i>
Step 1		.07***				.67***				.48***		
1st-grade academic functioning	.10**		0.20**	.06	.74***		0.79***	.02	.56***		0.56***	.03
Child gender	.09**		1.66**	.58	.02		0.39	.44	.07**		0.15**	.05
Child race	-.09*		-2.17**	.70	-.06**		-1.87**	.65	-.08**		-0.23***	.06
Maternal education	.00		0.02	.13	.12***		0.57***	.11	.14***		0.07***	.01
Family income-to-needs ratio	.12**		0.41***	.12	.04		0.14	.10	.09**		0.03**	.01
Maternal depression	-.09*		-0.10*	.04	.03		0.04	.04	-.01		-0.00	.00
Step 2		.01**				.00**				.01***		
Early insensitivity	-.13**		-1.01**	.34	-.08**		-0.95**	.30	-.11***		-0.12***	.03
Step 3		.01^				.01**				.02***		
Elementary school involvement	.09*		0.86**	.32	.07**		0.57*	.28	.17***		0.16***	.03
Elementary school insensitivity	-.04		-0.14	.36	-.06*		-0.73*	.29	-.09**		-0.07*	.03
Step 4		.01**				.01***				.00^		
Early Insensitivity $\times$ Elementary School Involvement	.10**		0.78**	.27	.08***		0.91***	.22	.05		0.02	.02

Note. For child gender, 1 = male and 2 = female; for child race, 1 = White and 2 = non-White.

^  $p < .10$ . \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

## Central Analyses

To evaluate whether the adverse effects of early insensitivity on children's later academic functioning can be offset by later involvement in children's education, hierarchical multiple regression was used. Each form of academic functioning (i.e., observations of classroom engagement, standardized achievement test scores, and teacher reports of academic competencies) at the end of elemen-

tary school (i.e., fifth grade) was predicted from early insensitivity (i.e., 6–54 months) and elementary school involvement (i.e., first to fifth grades). To take into account academic functioning in the initial school years, academic functioning at first grade was entered as a covariate in the first step. This step also included the other covariates: child gender (1 = male, 2 = female) and race (1 = White, 2 = non-White), maternal education (i.e., years of

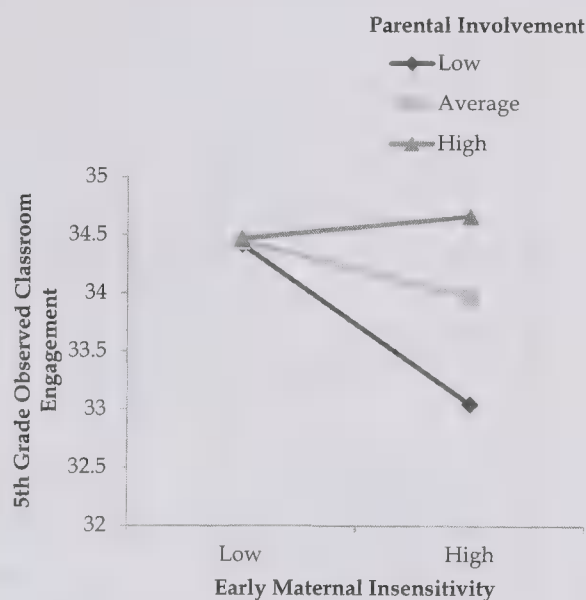


Figure 1. Parents' elementary school involvement moderates the effects of mothers' early insensitivity on observations of children's classroom engagement at fifth grade, controlling for engagement at first grade. Slopes were estimated from the regression equation from the original data set. Low insensitivity and involvement reflects estimates at the 25th percentile; average reflects estimates at the 50th percentile; high reflects estimates at the 75th percentile.

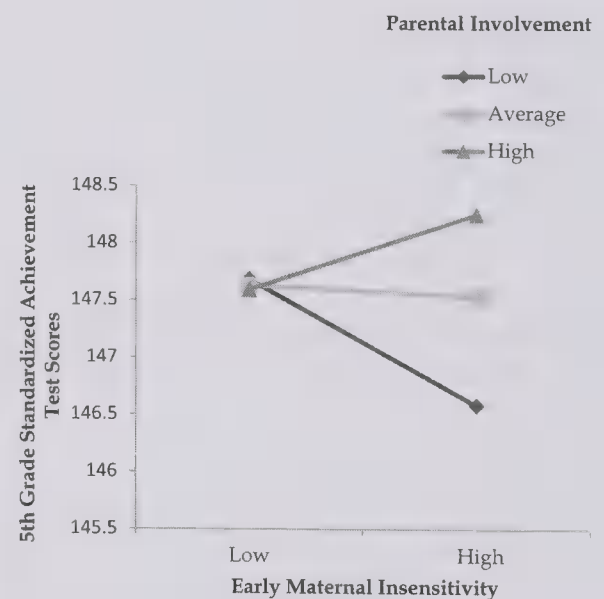


Figure 2. Parents' elementary school involvement moderates the effects of mothers' early insensitivity on children's scores on the Woodcock-Johnson standardized achievement tests at fifth grade, controlling for scores at first grade. Slopes were estimated from the regression equation from the original data set. Low insensitivity and involvement reflects estimates at the 25th percentile; average reflects estimates at the 50th percentile; high reflects estimates at the 75th percentile.

education completed) and depression, and the family income-to-needs ratio. In the second step, early insensitive parenting was entered. Elementary school involvement and insensitivity were added in the third step. This was followed by a fourth step including the target Early Insensitivity  $\times$  Elementary School Involvement interaction. As suggested by Aiken and West (1991), to reduce multicollinearity, the continuous variables were mean-centered by standardizing them.

As shown in Table 3, early insensitivity predicted academic functioning at the end of elementary school, taking into account such functioning at the beginning of elementary school: The more insensitive mothers were early on, the poorer children's subsequent engagement, standardized test scores, and academic competencies ( $t_s \geq 3.00$ ,  $p_s < .01$ ). It was also the case that the more involved parents were in children's education during elementary school, the better children's engagement, standardized test scores, and academic competencies at the end of elementary school ( $t_s \geq 2.12$ ,  $p_s < .05$ ). As anticipated, the effects of early insensitivity were moderated by elementary school involvement, as indicated by Early Insensitivity  $\times$  Elementary School Involvement interactions in predicting observations of engagement in the classroom ( $t = 2.90$ ,  $p < .01$ ) and standardized achievement test performance ( $t = 4.10$ ,  $p < .001$ ). The Early Insensitivity  $\times$  Elementary School Involvement interaction, however, did not reach significance for teachers' reports of academic competencies ( $t = 1.88$ ,  $p = .06$ ).

The interactions for observed classroom engagement and standardized achievement test performance were decomposed following Aiken and West's (1991) guidelines. To evaluate the effects of early insensitivity when elementary school involvement was low, we conducted hierarchical multiple regression analyses identical to those described earlier, but centering involvement at the 25th percentile (i.e., a standardized score of  $-.67$ ). The effects of early insensitivity when involvement was average or high were evaluated in regression analyses in which involvement was centered at the 50th (i.e., a standardized score of  $.16$ ) and 75th percentile (i.e., a standardized score of  $.79$ ), respectively. As shown in Figure 1, when elementary school involvement was low (i.e., the 25th percentile), early insensitivity was predictive of dampened classroom engagement at the end of elementary school ( $\beta = -.12$ ;  $t = 2.42$ ,  $p < .05$ ); however, when elementary school involvement was average (i.e., the 50th percentile) or high (i.e., the 75th percentile), early insensitivity did not matter for children's classroom engagement ( $\beta_s = -.04$  and  $.02$ ;  $t_s < 1$ ). Similarly, as shown in Figure 2, early insensitivity predicted lower achievement scores over time when elementary school involvement was low ( $\beta = -.07$ ;  $t = 2.44$ ,  $p < .05$ ), but not when it was average ( $\beta = -.01$ ;  $t = 0.22$ ,  $p = .83$ ) or high ( $\beta = .04$ ;  $t = 1.24$ ,  $p = .21$ ).

To identify the extent of involvement necessary to offset the adverse effects of early insensitivity, regions of significance analyses were conducted (see Aiken & West, 1991; Preacher, Curran, & Bauer, 2006; Roisman et al., 2012). The regions of significance analyses allowed us to identify the specific value of involvement (i.e., the moderator) at which early insensitivity no longer predicted dampened academic functioning. These analyses were conducted with the web-based application (<http://www.yourpersonality.net/interaction>) created by R. Chris Fraley as a supplement to Roisman et al. (2012). The results indicated that the minimum value of parents' involvement necessary to reduce the negative association between early insensitivity and observed classroom

engagement to nonsignificance was a standardized score of  $-.40$ , which is slightly less than a half standard deviation below the mean, representing the 33rd percentile of the sample. For standardized achievement test performance, the minimum value was a standardized score of  $-.47$ , which is almost a half standard deviation below the mean, representing the 32nd percentile of the sample.

To identify the proportion of children exposed to early insensitivity who were helped by parents' involvement in children's education, we identified the threshold of early insensitivity where parents' involvement began to have a significant positive effect—this is at slightly above the 50th percentile of early insensitivity. We then looked at how many families had scores that fell at or above this threshold and at or above the threshold of involvement that reduces the negative insensitivity effect to nonsignificance. The proportion of families ranged from 16% (for the analyses predicting engagement) to 22% (for the analyses predicting achievement)—that is, 179 to 241 families of those with the relevant data ( $n = 1,109$ ).

### Supplementary Analyses

To ensure that the effects we identified reflect the role of parents' involvement during elementary school rather than insensitive parenting during this time, we conducted hierarchical multiple regressions identical to those in the central analyses but replacing the Early Insensitivity  $\times$  Elementary School Involvement interaction with the Early Insensitivity  $\times$  Elementary School Insensitivity interaction. Although this interaction was not evident for observations of classroom engagement ( $\beta = -.02$ ;  $t = .65$ ,  $p = .52$ ), it was evident for standardized achievement test performance ( $\beta = -.08$ ;  $t = 3.75$ ,  $p < .001$ ) and teachers' reports of competencies ( $\beta = -.06$ ;  $t = 2.38$ ,  $p < .05$ ). To ensure that such an interaction was not responsible for the Early Insensitivity  $\times$  Elementary School Involvement interaction for standardized achievement test performance, the two interactions were entered simultaneously. Both the Early Insensitivity  $\times$  Elementary School Involvement interaction ( $\beta = .06$ ;  $t = 2.60$ ,  $p < .01$ ) and the Early Insensitivity  $\times$  Elementary School Insensitivity interaction ( $\beta = -.05$ ;  $t = 2.01$ ,  $p < .05$ ) remained. Thus, the compensatory role of elementary school involvement does not appear to be attributable to dampened insensitivity during elementary school.

Because prior research indicates that parents' involvement in children's education moderates the effect of parents' educational attainment on children's academic functioning, we wanted to ensure that elementary school involvement moderated the effects of early insensitivity over and above maternal educational attainment. Thus, regression analyses identical to those used in the central analyses were conducted replacing the Early Insensitivity  $\times$  Elementary School Involvement interaction with a Maternal Educational Attainment  $\times$  Elementary School Involvement interaction. Consistent with prior research, this interaction was evident for observed classroom engagement ( $\beta = -.07$ ;  $t = 2.08$ ,  $p < .05$ ) and standardized achievement test performance ( $\beta = -.08$ ;  $t = 3.99$ ,  $p < .001$ ); however, it did not reach significance for teacher reports of academic competencies ( $\beta = -.05$ ;  $t = 1.80$ ,  $p = .07$ ). To ensure that the Maternal Educational Attainment  $\times$  Elementary School Involvement interaction did not account for the Early Insensitivity  $\times$  Elementary School Involvement interaction, both



were simultaneously entered in the analyses, predicting observations of classroom engagement and standardized achievement test performance. The Early Insensitivity  $\times$  Elementary School Involvement interaction remained for engagement ( $\beta = .09$ ;  $t = 2.26$ ,  $p < .05$ ) and test scores ( $\beta = .06$ ;  $t = 2.64$ ,  $p < .01$ ). The Educational Attainment  $\times$  Elementary School Involvement interaction remained only for standardized achievement test scores ( $\beta = -.05$ ;  $t = 2.47$ ,  $p < .05$ ).

## Discussion

The current findings are consistent with the idea that parents' involvement in children's education can offset the adverse effects of early insensitive parenting on children's academic functioning. When parents were relatively uninvolved in children's education during elementary school (i.e., first to fifth grade), the legacy of early (i.e., 6–54 months) insensitive parenting was evident in deficits in children's classroom engagement and performance on standardized achievement tests at the end of elementary school (i.e., fifth grade). However, such effects were no longer statistically significant when parents showed average or higher involvement. Although parents' involvement in children's education is less common among parents with a history of insensitivity, examination of its compensatory role is important because it provides a window into its potential power to overcome academic risk among children.

### The Compensatory Role of Parents' Involvement in Children's Education

Replicating prior research indicating that early insensitive parenting foreshadows academic problems among children (e.g., Fralley et al., 2013; Lemelin et al., 2006; NICHD ECCRN, 2008; Stams et al., 2002), the current research revealed that the more insensitive mothers were early in children's lives (i.e., 6–54 months), the poorer children's academic functioning at the end of elementary school (i.e., fifth grade) over and above their earlier (i.e., first grade) academic functioning. However, this was moderated by parents' involvement in children's education during elementary school (i.e., first to fifth grades) such that mothers' early insensitivity predicted dampened subsequent academic functioning (i.e., fifth grade) among children only when parents were relatively uninvolved. Notably, it appeared that merely average (or better) involvement was necessary to offset the adverse effects of early insensitivity. Our analyses ruled out the possibility that the effects of parents' involvement simply reflected a change in insensitive parenting such that involved parents with a history of insensitive parenting became sensitive over time. However, remaining to be investigated is whether the compensatory role of parents' involvement in children's education begins earlier than the elementary school years: When insensitive parents are involved early in children's lives (e.g., by reading to them or counting with them), does this protect children? Given that our analyses took children's academic functioning at first grade into account, the compensatory role of such early involvement is likely distinct from of parents' involvement during elementary school.

The results of the current research parallel Dearing and colleagues' (Dearing, Kreider, Simpkins, & Weiss, 2006; Dearing, McCartney, Weiss, Kreider, & Simpkins, 2004) findings that par-

ents' involvement in children's education reduces achievement disparities between children of less versus highly educated mothers. Notably, the compensatory role of parents' involvement for early insensitive parenting is unique to insensitivity in that the Early Insensitivity  $\times$  Elementary School Involvement interaction remained when adjusting for the Educational Attainment  $\times$  Elementary School Involvement interaction. Taken together with Dearing and colleagues' findings, the current research suggests that parents' involvement may be powerful in its ability to offset multiple aspects of children's environment that put them at risk academically. The findings are also in line with Bronfenbrenner's (1992) ecological systems theory in underscoring that the influence of the microsystem (i.e., insensitive parenting) may be moderated by the connections parents establish outside this system (i.e., the mesosystem) as reflected in their involvement in children's education.

A major strength of the current research is the use of multiple methods to assess parenting and children's academic functioning. The two aspects of parenting examined were assessed using different methods—observations of mothers' insensitivity and teacher reports of parents' involvement—with multiple assessments over the time frames of interest. In addition, we examined three forms of children's academic functioning, each assessed with a different method (i.e., observations of classroom engagement, performance on standardized achievement tests, and teacher reports of academic competencies). Given that the compensatory effect of parents' involvement was evident for both observations of children's classroom engagement and children's performance on standardized tests, the effect's range of influence on academic functioning appears to be broad, with no evidence that it simply reflects reporter bias or shared method variance. In the case in which teachers reported on both parents' involvement and children's academic competencies, it may have been difficult to detect the target interaction because of the shared reporter bias. Indeed, parents' involvement was a larger predictor of this form of academic functioning than it was of observations of children's engagement or standardized achievement tests—a pattern that was not evident for insensitive parenting (see Table 3). Moreover, when we examined only the items asking about parents' involvement explicitly on the school front, the interaction predicting teachers' reports of children's academic competencies reached significance. The explicit nature of these items (vs. those that are more general) may minimize teachers' biases in completing the parent involvement measure.

At first blush, the finding that parents' involvement in children's education offsets the negative effects of early insensitive parenting may appear to contradict the argument made by several investigators that although the quantity of parents' involvement in children's education is important, so is the quality (e.g., Pomerantz, Grolnick, & Price, 2005; Pomerantz et al., 2007). Much research indicates that parents' sensitive involvement in children's education (e.g., supporting children's autonomy while helping them with an academic activity) is beneficial for children's academic functioning (e.g., Grolnick, Gurland, DeCoursey, & Jacob, 2002; Hokoda & Fincham, 1995; Moorman & Pomerantz, 2008). Two studies reveal an interaction between parents' involvement and what might be considered insensitive parenting opposite to the one identified here. Focusing on kindergarten students, Simpkins and colleagues (2006) found that parents' school-based involvement



was positively associated with children's achievement only when there was a warm parent-child relationship. Similarly, in a study with adolescents, parents' involvement on the school and home fronts was more predictive of achievement, but not necessarily engagement, the more parents used an authoritative parenting style (Steinberg et al., 1992).

These two studies examined parenting quality and parents' involvement in children's education at the *same* time; the current study, however, examined the two at *different* times—parenting quality in the years before children entered school and parents' involvement when children were in elementary school. In addition, in contrast to these two studies, recent research indicates that even when parents' involvement is accompanied by insensitive parenting (e.g., intrusiveness), it may benefit children in terms of their engagement and achievement in school, albeit not necessarily feelings of confidence and emotional functioning (Cheung & Pomerantz, 2011). Parents may also not need to be sensitive to become involved in children's education. In some cases, parents' involvement may be elicited by teachers (e.g., Green et al., 2007), rather than stemming from parents' sensitive concern for children; involvement may also not require substantial parent-child interaction. For example, teachers may bring children's academic problems to parents' attention and support parents in addressing such problems by referring them to resources such as tutoring services, or keeping parents up to date on children's progress. Parents' involvement on the school front may be particularly likely to confer benefits in the context of insensitive parenting because it requires relatively little interaction with children while still conveying the value of school and directing teachers' attention to children.

### Limitations and Future Directions

There are several limitations of the current research that suggest caution in interpreting the findings. Following much prior research (e.g., Englund, Luckner, Whaley, & Egeland, 2004; Grolnick & Slowiaczek, 1994; Izzo, Weissberg, Kasprow, & Fendrich, 1999), teachers' reports were used to assess parents' involvement in children's education. However, parents' involvement takes place at home as well as school. Although teachers may be able to accurately report on parents' involvement on the school front, they may not be able to do so when it comes to the home front, particularly for parents with whom they do not have much contact or insight into their cultural practices. Half of the items in the measure used in the current research ask about parents' involvement explicitly on the school front; the other half asked about it more generally. Future research should incorporate teachers, parents, and children's reports, with explicit assessment of parents' involvement on both the school and home fronts. We have suggested that parents' involvement on the school front may be particularly likely to compensate for early insensitivity because, unlike parents' involvement on the home front, it does not require substantial direct interaction with children. However, it is possible that parents' involvement on the home front also plays a compensatory role because it conveys the importance of school while also providing useful instruction and practices.

The assessment of parents' involvement in children's education used in this research did not distinguish between mothers' and fathers' involvement. Thus, it is unclear whether mothers, fathers,

or both are pivotal in offsetting the effects of mothers' early insensitivity on children's academic functioning. It could be that mothers who previously displayed insensitive parenting became involved in children's education during elementary school, thereby compensating for the effects of their own earlier insensitivity. It is also plausible that it was fathers who compensated for the effects; it may be that when children are struggling as a result of mothers' early insensitivity, fathers step in to provide support to children, either on their own or in conjunction with mothers. In line with this possibility, McBride, Dyer, Liu, Brown, and Hong (2009) suggested that fathers tend to become involved when children are having difficulties in school. Another limitation is that we examined parents' involvement over the whole of elementary school rather than at each year of this phase of development. This reflected our assumption that the effects of early insensitivity cannot be offset with a brief interlude of involvement, needing instead sustained involvement. However, future research using time-varying analyses could reveal whether short periods of involvement also play a compensatory role. A profile approach could also be taken to identify distinct profiles that take into account different levels of early insensitivity and involvement as well as their consistency over time.

A key question for future research is to what extent the compensatory role of parents' involvement in children's learning documented here extends to children's functioning beyond the academic arena. Exposure to early insensitivity has been implicated in the development of social and emotional problems among children (e.g., Fraley et al., 2013; Haltigan, Roisman, & Fraley, 2013; NICHD ECCRN, 2004; Raby et al., 2013; Stams et al., 2002). Given that parents' involvement in children's education also plays a role in social and emotional functioning (e.g., Cheung & Pomerantz, 2011; Hill et al., 2004), it may mitigate the effects of parents' early insensitivity on these aspects of functioning. However, parents' involvement in children's education may need to be accompanied by sensitive parenting, such as autonomy support, to ameliorate social and emotional problems (Cheung & Pomerantz, 2011). In addition, other parenting practices that directly target the psychological resources that facilitate social and emotional functioning may be more effective than parents' involvement in children's education—for example, parents' involvement in children's social lives as manifest in assisting children with developing strategies for resolving peer conflict.

### Conclusions

Decades of research on the role of the early environment in children's academic functioning have lead investigators to argue for its importance (e.g., Fraley et al., 2013; Heckman, 2006; Raby et al., 2013). The findings of the current research are consistent with this perspective in that early insensitive parenting predicted children's academic functioning at the end of elementary school. However, these effects were sizably reduced when parents were involved in children's education during the elementary school years. The results of the current research suggest that the detrimental effects of early insensitive parenting for academic functioning may not be unalterable; parents' involvement in children's education appears to be a potential avenue for helping children overcome academic problems created by their early environment. Thus, future research would benefit from identifying what enables



parents with a history of insensitive parenting to become involved in children's education.

## References

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Bornstein, M. H., & Tamis-LeMonda, C. S. (1997). Maternal responsiveness and infant mental abilities: Specific predictive relations. *Infant Behavior & Development*, 20, 283–296. doi:10.1016/S0163-6383(97)90001-1
- Bretherton, I. (1985). Attachment theory: Retrospect and prospect. *Monographs of the Society for Research in Child Development*, 50, 3–35. doi:10.2307/3333824
- Bronfenbrenner, U. (1992). *Ecological systems theory*. London, England: Jessica Kingsley.
- Campbell, S. B., Matestic, P., von Stauffenberg, C., Mohan, R., & Kirchner, T. (2007). Trajectories of maternal depressive symptoms, maternal sensitivity, and children's functioning at school entry. *Developmental Psychology*, 43, 1202–1215. doi:10.1037/0012-1649.43.5.1202
- Cheung, C. S.-S., & Pomerantz, E. M. (2011). Parents' involvement in children's learning in the United States and China: Implications for children's academic and emotional adjustment. *Child Development*, 82, 932–950. doi:10.1111/j.1467-8624.2011.01582.x
- Dearing, E., Kreider, H., Simpkins, S., & Weiss, H. B. (2006). Family involvement in school and low-income children's literacy: Longitudinal associations between and within families. *Journal of Educational Psychology*, 98, 653–664. doi:10.1037/0022-0663.98.4.653
- Dearing, E., McCartney, K., Weiss, H. B., Kreider, H., & Simpkins, S. (2004). The promotive effects of family educational involvement for low-income children's literacy. *Journal of School Psychology*, 42, 445–460. doi:10.1016/j.jsp.2004.07.002
- De Wolff, M., & van IJzendoorn, M. (1997). Sensitivity and attachment: A meta-analysis on parental antecedents of infant attachment. *Child Development*, 68, 571–591. doi:10.2307/1132107
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.
- Englund, M. M., Luckner, A. E., Whaley, G. J. L., & Egeland, B. (2004). Children's achievement in early elementary school: Longitudinal effects of parental involvement, expectations, and quality of assistance. *Journal of Educational Psychology*, 96, 723–730. doi:10.1037/0022-0663.96.4.723
- Epstein, J. L. (1988). How do we improve programs for parental involvement? *Educational Horizons*, 66, 75–77.
- Epstein, J. L., & Becker, H. J. (1982). Teachers' reported practices of parent involvement: Problems and possibilities. *Elementary School Journal*, 83, 103–113. doi:10.1086/461298
- Fraley, R. C., Roisman, G. I., & Haltigan, J. D. (2013). The legacy of early experiences in development: Formalizing alternative models of how early experiences are carried forward over time. *Developmental Psychology*, 49, 109–126. doi:10.1037/a0027852
- Frodi, A., Bridges, L., & Grolnick, W. S. (1985). Correlates of mastery-related behavior: A short-term longitudinal study of infants in their second year. *Child Development*, 56, 1291–1298. doi:10.2307/1130244
- Green, C. L., Walker, J. M. T., Hoover-Dempsey, K. V., & Sandler, H. M. (2007). Parents' motivations for involvement in children's education: An empirical test of a theoretical model of parental involvement. *Journal of Educational Psychology*, 99, 532–544. doi:10.1037/0022-0663.99.3.532
- Grolnick, W. S., Gurland, S. T., DeCoursey, W., & Jacob, K. (2002). Antecedents and consequences of mothers' autonomy support: An experimental investigation. *Developmental Psychology*, 38, 143–155. doi:10.1037/0012-1649.38.1.143
- Grolnick, W. S., & Slowiaczek, M. L. (1994). Parents' involvement in children's schooling: A multidimensional conceptualization and motivational model. *Child Development*, 65, 237–252. doi:10.2307/1131378
- Haltigan, J. D., Roisman, G. I., & Fraley, R. C. (2013). The predictive significance of early caregiving experiences for symptoms of psychopathology through mid-adolescence: Enduring or transient effects? *Development and Psychopathology*, 25, 209–221. doi:10.1017/S0954579412000260
- Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, 312, 1900–1902. doi:10.1126/science.1128898
- Hill, N. E., Castellino, D. R., Lansford, J. E., Nowlin, P., Dodge, K. A., Bates, J. E., & Pettit, G. S. (2004). Parent academic involvement as related to school behavior, achievement, and aspirations: Demographic variations across adolescence. *Child Development*, 75, 1491–1509. doi:10.1111/j.1467-8624.2004.00753.x
- Hokoda, A., & Fincham, F. D. (1995). Origins of children's helpless and mastery achievement patterns in the family. *Journal of Educational Psychology*, 87, 375–385. doi:10.1037/0022-0663.87.3.375
- IBM Corp. (2010). *IBM SPSS Statistics for Windows, Version 19.0*. Armonk, NY: Author.
- Izzo, C. V., Weissberg, R. P., Kasprow, W. J., & Fendrich, M. (1999). A longitudinal assessment of teacher perceptions of parent involvement in children's education and school performance. *American Journal of Community Psychology*, 27, 817–839. doi:10.1023/A:1022262625984
- Kohl, G. O., Lengua, L. J., McMahon, R. J., & Conduct Problems Prevention Research Group. (2000). Parent involvement in school: Conceptualizing multiple dimensions and their relations with family and demographic risk factors. *Journal of School Psychology*, 38, 501–523. doi:10.1016/S0022-4405(00)00050-9
- Lemelin, J.-P., Tarabulsy, G. M., & Provost, M. (2006). Predicting pre-school cognitive development from infant temperament, maternal sensitivity, and psychosocial risk. *Merrill-Palmer Quarterly*, 52, 779–804. doi:10.1353/mpq.2006.0038
- Linver, M. R., Brooks-Gunn, J., & Kohen, D. E. (2002). Family processes as pathways from income to young children's development. *Developmental Psychology*, 38, 719–734. doi:10.1037/0012-1649.38.5.719
- Main, M. (1983). Exploration, play, and cognitive functioning related to infant-mother attachment. *Infant Behavior and Development*, 6, 167–174. doi:10.1016/S0163-6383(83)80024-1
- Matas, L., Arend, R. A., & Sroufe, L. A. (1978). Continuity of adaptation in the second year: The relationship between quality of attachment and later competence. *Child Development*, 49, 547–556. doi:10.2307/1128221
- McBride, B. A., Dyer, W. J., Liu, Y., Brown, G. L., & Hong, S. (2009). The differential impact of early father and mother involvement on later student achievement. *Journal of Educational Psychology*, 101, 498–508. doi:10.1037/a0014238
- Miller-Johnson, S., & Maumary-Gremaud, A., & Conduct Disorders Research Group. (1995). *Parent-Teacher Involvement: Teacher version*. Durham, NC: Duke University.
- Moorman, E. A., & Pomerantz, E. M. (2008). Mothers' cognitions about children's self-control: Implications for mothers' responses to children's helplessness. *Social Development*, 17, 960–979. doi:10.1111/j.1467-9507.2008.00469.x
- National Institute of Child Health and Human Development Early Child-care Research Network. (1997). The effects of infant child care on infant-mother attachment security. *Child Development*, 68, 860–879. doi:10.1111/j.1467-8624.1997.tb01967.x
- National Institute of Child Health and Human Development Early Child-care Research Network. (2004). Fathers' and mothers' parenting behavior and beliefs as predictors of children's social adjustment in the transition to school. *Journal of Family Psychology*, 18, 628–638. doi:10.1037/0893-3200.18.4.628
- National Institute of Child Health and Human Development Early Child-care Research Network. (2005). Duration and developmental timing of poverty and children's cognitive and social development from birth

- through third grade. *Child Development*, 76, 795–810. doi:10.1111/j.1467-8624.2005.00878.x
- National Institute of Child Health and Human Development Early Child-care Research Network. (2008). Mothers' and fathers' support for child autonomy and early school achievement. *Developmental Psychology*, 44, 895–907. doi:10.1037/0012-1649.44.4.895
- Nicholson, J., Atkins-Burnett, S., & Meisels, S. (n.d.). *ECLS-K base year public-use data files and electronic codebook*. Retrieved from <http://nces.ed.gov/pubs2001/2001029rev.pdf>
- Nolen-Hoeksema, S., Wolfson, A., Mumme, D., & Guskin, K. (1995). Helplessness in children of depressed and nondepressed mothers. *Developmental Psychology*, 31, 377–387. doi:10.1037/0012-1649.31.3.377
- Pierce, K. M., Hamm, J. V., & Vandell, D. L. (1999). Experiences in after-school programs and children's adjustment in first-grade classrooms. *Child Development*, 70, 756–767. doi:10.1111/1467-8624.00054
- Pomerantz, E. M., Grolnick, W. S., & Price, C. A. (2005). The role of parents in how children approach achievement: A dynamic process perspective. In A. Elliot & C. W. Dweck (Eds.), *Handbook of competence and motivation* (pp. 259–278). New York, NY: Guilford Press.
- Pomerantz, E. M., & Moorman, E. A. (2010). Parents' involvement in children's school lives: A context for children's development. In J. Meece & J. Eccles (Eds.), *Handbook of research on schools, schooling, and human development* (pp. 398–416). New York, NY: Routledge.
- Pomerantz, E. M., Moorman, E. A., & Cheung, C. S. (2012). Parents' involvement in children's learning. In K. R. Harris, S. Graham, T. C. Urdan, S. Graham, J. M. Royer, & M. Zeidner (Eds.), *APA educational psychology handbook* (pp. 417–440). Washington, DC: American Psychological Association.
- Pomerantz, E. M., Moorman, E. A., & Litwack, S. D. (2007). The how, whom, and why of parents' involvement in children's schooling: More is not necessarily better. *Review of Educational Research*, 77, 373–410. doi:10.3102/003465430305567
- Pomerantz, E. M., Wang, Q., & Ng, F. F. (2005). Mothers' affect in the homework context: The importance of staying positive. *Developmental Psychology*, 41, 414–427. doi:10.1037/0012-1649.41.2.414
- Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, 31, 437–448. doi:10.3102/10769986031004437
- Raby, K. L., Roisman, G. I., Fraley, R. C., & Simpson, J. A. (2013). *The predictive significance of early maternal sensitivity: Academic and social competence through age 32 years in the Minnesota Longitudinal Study of Risk and Adaptation*. Manuscript submitted for publication.
- Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1, 385–401. doi:10.1177/014662167700100306
- Riksen-Walraven, J. M. (1978). Effects of caregiver behavior on habituation rate and self-efficacy in infants. *International Journal of Behavioral Development*, 1, 105–130. doi:10.1177/016502547800100202
- Roisman, G. I., & Fraley, R. C. (2012). A behavior-genetic study of the legacy of early caregiving experiences: Academic skills, social competence, and externalizing behavior in kindergarten. *Child Development*, 83, 728–742. doi:10.1111/j.1467-8624.2011.01709.x
- Roisman, G. I., Newman, D. A., Fraley, R. C., Haltigan, J. D., Groh, A. M., & Haydon, K. C. (2012). Distinguishing differential susceptibility from diathesis stress: Recommendations for evaluating interaction effects. *Development and Psychopathology*, 24, 389–409. doi:10.1017/S0954579412000065
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley. doi:10.1002/9780470316696
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177.
- Sénéchal, M., & LeFevre, J. (2002). Parental involvement in the development of children's reading skill: A five-year longitudinal study. *Child Development*, 73, 445–460. doi:10.1111/1467-8624.00417
- Simpkins, S. D., Weiss, H. B., McCartney, K., Kreider, H. M., & Dearing, E. (2006). Mother-child relationship as a moderator of the relation between family educational involvement and child achievement. *Parenting: Science and Practice*, 6, 49–57. doi:10.1207/s15327922par0601\_2
- Stams, G. J. M., Juffer, F., & van IJzendoorn, M. (2002). Maternal sensitivity, infant attachment, and temperament in early childhood predict adjustment in middle childhood: The case of adopted children and their biologically unrelated parents. *Developmental Psychology*, 38, 806–821. doi:10.1037/0012-1649.38.5.806
- Steinberg, L., Lamborn, S. D., Dornbusch, S. M., & Darling, N. (1992). Impact of parenting practices on adolescent achievement: Authoritative parenting, school involvement, and encouragement. *Child Development*, 63, 1266–1281. doi:10.2307/1131532
- van IJzendoorn, M. H., Dijkstra, J., & Bus, A. G. (1995). Attachment, intelligence, and language: A meta-analysis. *Social Development*, 4, 115–128. doi:10.1111/j.1467-9507.1995.tb00055.x
- Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery-Revised*. Allen, TX: DLM.

Received August 23, 2013

Revision received December 17, 2013

Accepted December 29, 2013 ■



# Strengthening Bullying Prevention Through School Staff Connectedness

Lindsey M. O'Brennan, Tracy E. Waasdorp, and Catherine P. Bradshaw  
Johns Hopkins University

The growing concern about bullying and school violence has focused national attention on various aspects of school climate and school connectedness. The current study examined dimensions of staff connectedness (i.e., personal, student, staff, and administration) in relation to staff members' comfort intervening in bullying situations (e.g., physical, verbal, relational), as well as bullying situations involving special populations of students (e.g., gender-nonconforming, disability, overweight, sexism, racism, and religion). Data for this study were collected from a national sample of 5,064 members of the National Education Association (NEA), of whom 2,163 were teachers and 2,901 other school staff. Analyses with structural equation modeling indicated that increased staff connectedness was associated with greater comfort intervening with bullying. Similarly, having resources available regarding bullying, receiving training on the school's bullying policy, and being involved in bullying prevention efforts were significantly associated with comfort intervening. Implications for school-based prevention and school climate promoting efforts are discussed.

**Keywords:** bullying intervention, school connectedness, school staff development, youth violence prevention

Many prevention and intervention programs have the dual focus of reducing children's aggressive and violent behaviors while promoting caring and supportive environments for students and staff (Gilman, Huebner, & Furlong, 2009; Thapa, Cohen, Guffey, & Higgins-D'Alessandro, 2013). Relationships among individuals in the school environment are key contributors to student and staff perceptions of their school, often referred to as *school connectedness* (Libbey, 2004). Several youth violence prevention programs emphasize connectedness among students, teachers, administrators, and educational support professionals as a way of increasing staff buy-in and program effectiveness (Beets et al., 2008; Bradshaw, Koth, Thornton, & Leaf, 2009; Greenberg et al., 2003). A number of studies have highlighted the importance of teachers' and staff members' perceptions of the school (e.g., schools' organizational health) for high work productivity, staff efficacy, and focus on student success (Bevans, Bradshaw, Miech, & Leaf, 2007; Hoy & Woolfolk, 1993; Pas, Bradshaw, Hershfeldt, & Leaf, 2010). In particular, bullying prevention programs have increasingly focused on building positive relationships within the school

community as a means of reducing peer victimization and shifting schoolwide norms related to violence (Doll, Song, & Siemers, 2004; Olweus, Limber, & Mihalic, 1999).

The current study builds upon the school climate literature by examining multiple dimensions of school staff connectedness (i.e., staff–student connectedness, personal connectedness, staff–staff connectedness, and staff–administration connectedness) in conjunction with staff members' perceptions of bullying prevention programming efforts, as they relate to their willingness to intervene in bullying situations. This issue is particularly relevant for schools across the United States, given the nearly ubiquitous nature of school bullying and the associated academic, behavioral, and social–emotional risks for both perpetrators and targets (Swearer, Espelage, Vaillancourt, & Hymel, 2010). Therefore, it is critical that educators better understand factors that contribute to staff members' willingness to intervene in different types of bullying situations.

## Bullying Among School-Aged Youth

Bullying is broadly defined as intentional and repeated acts that occur through direct verbal (e.g., threatening, name calling), direct physical (e.g., hitting, kicking), and indirect (e.g., spreading rumors, influencing relationships, cyberbullying) forms, and it typically occurs in situations in which there is a power or status difference (Olweus, 1993). In a recent national survey, 75% of teachers had a student report a verbal bullying incident to them, 58% heard reports of relational bullying, 50% physical bullying, and 14% cyberbullying (Bradshaw, Waasdorp, & O'Brennan, 2010; Bradshaw, Waasdorp, O'Brennan, & Gulemetova, 2013). Although bullying affects roughly 30% of school-age youth (Bradshaw, Sawyer, & O'Brennan, 2007; Nansel et al., 2001), there are populations of students who are at an increased risk for peer victimization. For instance, research indicates that youth identifying as lesbian, gay, bisexual, or transgender (LGBT) are more

---

This article was published Online First February 17, 2014.

Lindsey M. O'Brennan, Tracy E. Waasdorp, and Catherine P. Bradshaw, Bloomberg School of Public Health, Johns Hopkins University.

The research reported here was supported in part by the National Education Association through a contract awarded to Catherine P. Bradshaw. The opinions expressed are those of the authors and do not represent views of the National Education Association. The authors thank Joann Morris and Michaela Gulemetova of the National Education Association for their support of this work. Additional support for the writing of this article came from the National Institute of Mental Health Children's Mental Health Services Training Program (T32 MH019545-21).

Correspondence concerning this article should be addressed to Lindsey M. O'Brennan, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205. E-mail: Lobrenna@jhsph.edu

likely to be targets of bullying than their heterosexual peers (Kosciw, Greytak, Diaz, & Bartkiewicz, 2010), which in turn increases their risk for psychological functioning in young adulthood (Russell, Ryan, Toomey, Diaz, & Sanchez, 2011). Students and staff members also report bullying related to weight to be a significant problem across grade levels (Bradshaw et al., 2013), with overweight or obese students being more likely to experience bullying than their peers (Brixval, Rayce, Rasmussen, Holstein, & Due, 2012). Likewise, students with disabilities, especially those lacking age-appropriate social skills and displaying behavior problems, are at an increased risk of being targets of peer victimization (Rose, Monda-Amaya, & Espelage, 2011; Zablotsky, Bradshaw, Anderson, & Law, 2012). Last, racial and ethnic minorities may also be more likely to report experiencing bullying at school (Sawyer, Bradshaw, & O'Brennan, 2008). Despite research consistently showing that special populations are frequent targets of bullying, little is known about the relation between school staff members' comfort intervening in bullying situations and their sense of connectedness to the school.

### **School Staff Connectedness and Bullying Prevention**

Accumulating evidence suggests school climate and school connectedness are multidimensional constructs that include school safety, quality of relationships, discipline practices, and aspects of the physical environment (Thapa et al., 2013; You, O'Malley, & Furlong, 2013; Zullig, Koopman, Patton, & Ubbes, 2010). Consistent with ecological systems theory (Bronfenbrenner, 1979; Bronfenbrenner & Morris, 1998), these evolving relationships play a role in staff members' sense of connectedness with others at school. In turn, this sense of affiliation and commitment with the school organization is expected to positively impact one's likelihood of intervening in bullying situations. The sections that follow summarize literature on four interrelated dimensions of staff connectedness: (a) personal sense of safety and connectedness to school, (b) student-staff relationships, (c) staff relationships to fellow employees, and (d) staff connectedness to administrators as they relate to bullying interventions across populations of students.

### **Personal Connectedness**

Personal connectedness is often thought of as a composite of feelings of respect and support from others at the school, perceptions of safety, and overall job satisfaction (Butler, 2012; Parker, Martin, Colmar, & Liem, 2012; Skaalvik & Skaalvik, 2011). An often-overlooked aspect of staff's personal connectedness is their perceived level of safety of the school environment. Although students' reports of safety are typically the catalyst for schoolwide violence prevention programs, national data revealed 7% of educators report being threatened and 4% of school staff report being physically attacked by a student (Keigher, 2009). Examining this in the context of schoolwide bullying efforts, it seems plausible that teachers and educational support professionals would be less likely to intervene in bullying situations when they perceive aggressive behavior to be the norm for the school (Kochenderfer-Ladd, & Pelletier, 2008). Conversely, when staff discern there to be a positive and prosocial climate at the school, they may feel more comfortable addressing issues of peer victimization, particularly in situations involving sensitive issues like ethnicity, obesity, and gender nonconformity.

### **Student-Staff Connectedness**

A long line of research has documented the importance of student-teacher relationships across grade levels. Student-teacher connectedness has been found to serve as a protective factor from the deleterious effects of bullying on students' academic achievement (Konishi, Hymel, Zumbo, & Li, 2010). Likewise, youth who report low school connectedness tend to report more instances of physical, verbal, and relational forms of peer victimization (O'Brennan & Furlong, 2010). From the school staff perspective, teachers' relationships with students are a strong predictor of their professional commitment to teaching and loyalty to their specific school (Collie, Shapka, & Perry, 2011). Teachers who are close with their students are more likely to report greater job satisfaction and teacher efficacy and reduced student problem behavior in the classroom (Collie, Shapka, & Perry, 2012). With regard to bullying prevention efforts, multilevel studies suggest schools with more positive student-teacher relationships tend to have reduced rates of bullying episodes (Richard, Schneider, & Mallet, 2012). Students may also be more likely to report bullying incidents to staff members with whom they have existing relationships, thus increasing the probability school staff will intervene. Furthermore, staff who feel personally connected to their students may be more likely to broach sensitive topics and more directly address topics that have been historically taboo in schools, such as ethnic differences and sexual orientation. Similarly, staff who feel personally connected to their students may also be inclined to put issues of difference aside and support students who are different than themselves.

### **Staff-Administration Connectedness**

School staff members' relationships with administrators also have been shown to be important, especially as they relate to implementing schoolwide programs and new initiatives. For example, program implementation research shows that it takes schools roughly 3–5 years to implement schoolwide programs with fidelity (Bradshaw, Reinke, Brown, Bevans, & Leaf, 2008); thus, it is essential for administrators to foster staff buy-in for program success. Strong working relationships among staff and administration are often forged through shared leadership on schoolwide policies and interventions. For instance, Sun, Shek, and Siu (2008) found that a key component to successful program implementation was teachers' ability to forge caring, respectful, and supportive relationships with the school administration. By forming positive staff-administrator relationships, schools are modeling positive interpersonal behaviors for students. Consequently, it was predicted that school staff who report positive relationships with their administrators would be more invested in school violence programming and in turn feel more comfortable addressing bullying. We further anticipated that administrator support would be especially important for addressing bullying situations among special populations due to the sensitive nature of these incidents.

### **Staff-Staff Connectedness**

Similar to staff-administrator relationships, school staff members' relationships with each other are salient aspects of school connectedness and program implementation. Teachers who openly



communicate with their peers also tend to be more open to professional growth and innovation (Collie et al., 2011). These findings have been replicated among schools most at risk for teacher attrition (Brown & Medway, 2007). In terms of bullying intervention, Kallestad and Olweus (2003) found that staff members' openness and communication with one another significantly impacted the implementation of an antibullying program. This association has been found to endure over time, with research showing that when teachers felt supported by their peers and administrators, they perceived the school climate more positively and delivered more lessons in a prevention curriculum (Gregory, Henry, Schoeny, & Metropolitan Area Study Research Group, 2007). Thus, it would appear that staff who feel connected to one another would be more invested and comfortable intervening in bullying at their school.

### Overview of the Current Study

In an effort to address gaps in bullying prevention literature, the current study explored four key aspects of staff connectedness: (a) personal connectedness to school, (b) student-staff relationships, (c) staff connectedness to administrators, and (d) staff relationships to fellow employees as they relate to comfort intervening with bullying in general, and more specifically in situations involving special populations of students. Structural equation modeling (SEM) was used to examine the association between school staff perceptions of connectedness and their comfort intervening with general bullying situations (physical, relational, verbal, and cyber) and when the bullying was specifically related to special populations (i.e., LGBT youth, students with disabilities, racial/ethnic minority students, youth who are overweight). An SEM approach was selected because it allowed us to test our hypotheses using a latent variable framework (Bollen, 1989; Kline, 2005). Based on the available research, we predicted that school staff who report higher levels of connectedness would be more likely to intervene in bullying compared with school staff reporting low levels of connectedness. Specifically, we hypothesized that particular forms of connectedness, including personal sense of safety and connectedness to school, student-staff relationships, and staff connectedness to administrators and their peers, would be associated with a greater likelihood of intervening in bullying among special populations. Finally, we examined whether the existence of formal bullying prevention programs and policies, involvement and training related to programming efforts, and staff perceptions of available bullying resources are associated with comfort intervening. We hypothesized that staff who report clear policies and prevention efforts had access to available resources and were involved in the training efforts would be more comfortable intervening.

### Method

#### Sample

Data for the current study come from the National Education Association (NEA), the country's largest teachers' union, which includes 3.2 million members nationwide. The authors partnered with the NEA to conduct a large-scale national study examining staff members' perceptions of bullying and the school environment. The sample included 5,064 adults who were members of the

NEA at the time of the data collection and were actively employed by a school or school system. A little over half of the sample was education support professionals (ESPs;  $n = 2,901$ ) and the remaining participants were teachers ( $n = 2,163$ ). As later described in greater detail, the sample was weighted to be representative of the full population of NEA members. Nearly half of the ESps were paraprofessionals (49%), followed by maintenance (14%), clerical (10%), school transportation (10%), food service (7%), health and student services (2%), technical and skilled trades (2%), security (1%), and other nonteaching support staff (6%). Women composed 80% of the sample, and 89% of the sample self-identified as White, with 5% Black, 4% Hispanic, and 2% other. The participants were employed in a variety of school locations (34% suburban, 24% small town, 24% urban, and 18% rural areas). Approximately 39% worked with students in elementary, 19% in middle, and 27% in high schools, with the remaining 16% working across multiple grade levels (see Watts, 2010).

### Procedure

The data were collected in the spring of 2010. In an effort to survey a representative sample of NEA members, both a telephone (63%) and a Web survey (37%) were used. Specifically, we used the Web because of growing concerns that individuals are less inclined to participate in and/or be reached by phone surveys (Holbrook, Krosnick, & Pfent, 2007). In total, 1,601 teachers and 2,142 ESps completed the telephone survey, whereas 562 teachers and 759 ESps completed the Web survey. The data collection activities were conducted by an external professional research firm contracted by the NEA; the subcontractor made the phone calls and administered the survey on behalf of the NEA. With regard to incentives for participation, a lottery was used, whereby all participants were informed that 20 participants would be selected at random for \$100. Participants were told that the purpose of the study was to inform the NEA about members' concerns and needs related to bullying and school climate. A sampling procedure was used to select participants, which accounted for role and select demographics (e.g., age, region, race), thereby allowing the data to be weighted up to reflect the entire population of NEA members; weighting is possible because of the known population distributions in the overall NEA membership database. As later described in greater detail, two weighting procedures were utilized on the data: a propensity score was used to adjust for the mode of survey administration (i.e., Web vs. phone) and a rim weight to weight the entire data set to the national population of NEA members (Watts, 2010). The overall participation rate was 31% (35% phone, 24% Web).

### Measure

The NEA Bullying Survey (see Bradshaw et al., 2010, for the complete survey; also see Bradshaw et al., 2013) was developed by the research team in close collaboration with the NEA Research Department. Consistent with previous studies of bullying (Bradshaw et al., 2007; Nansel et al., 2001; Olweus, 1993), bullying was defined on the survey as "intentional and repeated aggressive acts that can be physical (such as hitting); verbal (such as threats or name calling), or relational (such as spreading rumors or influencing social relationships). Bullying typically occurs in situations where there is a power or status difference."

**Connectedness.** Staff connectedness was assessed through items from the Charles F. Kettering Climate Scale (Johnson, Johnson, Kranch, & Zimmerman, 1999) and the Collegial Leadership subscale of the Organizational Health Inventory (Hoy & Woolfolk, 1993). In total, there were 21 items (full scale,  $\alpha = .95$ ) broken into four subscales: Personal Connectedness (eight items,  $\alpha = .89$ ; “My ideas are listened to; people care about me at this school,” “I like to work in my school”), student–staff connectedness (four items,  $\alpha = .90$ ; “Staff really care about the students”; “Staff are on students’ side”), staff–staff connectedness (five items,  $\alpha = .91$ ; “Staff are friendly to each other,” “Staff have trust and confidence in each other”), and staff administration connectedness (four items,  $\alpha = .90$ ; “Principal shows staff appreciation,” “Principal looks out for staff”). Response options were on a 4-point Likert scale from *disagree strongly* (1) to *agree strongly* (4), with higher scores indicating higher levels of connectedness.

**Comfort intervening with bullying.** Staff were asked how comfortable would they feel intervening with a student who engaged in bullying (4-point scale, ranging from *very uncomfortable* to *very comfortable*) across four forms of bullying (i.e., physical, relational, verbal, and cyber). The five forms of bullying were correlated (range .50–.74) and therefore were modeled as a single latent variable, “comfort intervening with general bullying” ( $\alpha = .87$ ).

**Comfort intervening with special populations.** Staff members were asked how comfortable would they feel intervening with a student who engaged in bullying (4-point scale, ranging from *very uncomfortable* to *very comfortable*) across six situations of bullying (i.e., bullying related to sexual orientation or gender nonconformity, disability, being overweight, sexism, racism, and religion). The six types of bullying were highly correlated (range .65–.84), and, therefore, were modeled as a single latent variable, “comfort intervening with special populations” ( $\alpha = .95$ ).

**Perceptions of bullying policies and programming.** Perceptions of the school’s bullying policies and programming were assessed through three yes/no questions: (a) Does your school district have a bullying policy? (b) Is the policy clear and easy to implement? (c) Did you receive training on how to implement the policy? Perceptions of bullying prevention efforts were assessed through two yes/no items: (a) Does the school you work in most frequently have a formal prevention efforts—such as school teams, a committee, or prevention program that deals with bullying? (b) Are you currently involved in bullying prevention activities at the school you work in most frequently? To assess resource availability, staff rated one item (“There are resources available to me to help me intervene with bullying”) on a 4-point scale from *disagree strongly* to *agree strongly*. These items were adapted from the measure by Bradshaw et al. (2007) (also see Bradshaw et al., 2013). All items had “Not sure” as an option, which was coded as missing in these analyses. The survey did employ a skip pattern, whereby if a staff member reported “no” to particular question, he or she would not be asked follow-up questions regarding that particular issue. For example, in the situation where a participant indicated that his or her school did not have a bullying prevention policy, that person was not asked follow-up questions regarding how easy it was to implement the policy. As a result, those individuals not asked a particular question were excluded from analyses of that question.

## Overview of Analyses

A series of SEMs was fit in to test our primary research questions related to the association between various aspects of school climate and staff members’ willingness to intervene in different bullying situations (i.e., general forms of bullying and with special populations). Based on prior research with this sample (Bradshaw et al., 2013), we adjusted for a set of covariates including amount of interaction between students and staff, school level (elementary and high school, with middle school as the reference category), school location (urban vs. suburban/rural), survey modality (Web vs. phone), and role in school (ESP vs. teacher). Staff age and number of years working in education were also included as continuous variables. We also applied sampling weights in all analyses (see later description). Missing data were generally not a concern, as 93% of the sample had no missing data, and each item had less than 2% missing. Utilizing Mplus Version 7.1 (Muthén & Muthén, 1998–2012), missing data are assumed to be missing at random and all analyses adjust for missing data using full information maximum likelihood.

**Sample weighting.** Two types of weights were applied to the data. First, we applied a propensity score weight to adjust for the mode of survey administration (i.e., Web vs. phone; Rosenbaum & Rubin, 1983; Schonlau, van Soest, Kapteyn, & Couper, 2009). The purpose of the propensity score weights was to make the Web-based survey comparable to the phone-based survey. Each participant was assigned a weight based on his or her propensity score, which was constructed based on 16 different demographic variables (e.g., full- vs. part-time worker, region of the country, has phone/landline, suburban location, years worked in the school, interaction level with students). These methods are commonly used in large-scale surveys that employ both phone and Web-based assessments (for additional details, see Schonlau et al., 2009; Taylor, 2000). Our decision to apply this type of weight was based on preliminary analyses of the data, which suggested that there were some systematic differences in the responses to select survey items based on the mode of survey administration. For example, phone respondents had a tendency to report greater comfort intervening in the different types of bullying situations assessed; this is likely due to a social desirability bias among phone participants (Kreuter, Presser, & Tourangeau, 2008; Watts, 2010). As a result, the propensity score weights, along with controlling for survey administration as a covariate in the analyses, allowed us to account for potential bias associated with those respondents who completed the Web survey compared with those who completed the phone survey. The second weight applied was a rim weight, which is a common weighting approach that enabled us to weight the entire data set to the national population of NEA members (Watts, 2010). Specifically, rim weighting was utilized to weight the sample that participated in the survey to those in the known NEA population. Therefore, the weighted sample reflects the full NEA membership.

## Results

### Sample Demographics

On average, staff reported working at the school for 10 years ( $SD = 8.85$ , range: 0–77 years) and ranged in age from 19 to 80



years old ( $M = 46.21$ ,  $SD = 14.55$ ). Majority of staff (74.4%) reported interacting with students constantly, 18.6% reported interacting with students "a great deal," and 7% reported very little interaction with students (i.e., "only a little" or "almost none"). Approximately 34% reported the school they work in was located in a suburban community, 23.9% an urban community, 24.1% a small town, and 17.6% in a rural community. Roughly 40% of staff worked in elementary schools, 33.4% in middle schools, and 27.3% in high schools (for additional descriptives, see Bradshaw et al., 2013).

### Fit of the Measurement Model

Confirmatory factor analysis was used to assess the fit of the four latent connectedness variables (i.e., personal connectedness, staff–staff connectedness, principal connectedness, and staff–student connectedness; see Table 1). This model demonstrated acceptable fit (comparative fit index [CFI] = .95, Tucker–Lewis index [TLI] = .94, root-mean square error of approximation [RMSEA] = .03, standardized root-mean-square residual [SRMR] =

0.04). Similarly, the latent variable "comfort intervening with general bullying," composed of the four different forms of bullying (i.e., physical, relational, verbal, and cyber) had adequate fit, CFI = .99, TLI = .96, RMSEA = .06, SRMR = .02 (see Table 1). Finally, the latent variable "comfort intervening with special populations," composed of the six different types of bullying situations (i.e., bullying related to sexual orientation/gender-nonconformity, disability, being overweight, sexism, racism, and religion), had adequate fit, CFI = .97, TLI = .95, RMSEA = .06, SRMR = .03 (see Table 1).

### Fit of the Structural Model

To assess our primary research aims, we fit a series of three SEM models for the two separate outcomes: (a) staff comfort intervening with general bullying and (b) staff comfort intervening with special populations.

**Staff connectedness.** Model 1 examined the relationship between the four connectedness latent variables (i.e., personal connectedness and student–staff, staff–administration, and staff–staff connectedness) and comfort intervening with bullying. For general bullying situations, Model 1 demonstrated acceptable fit, CFI = .95, TLI = .95, RMSEA = .03, SRMR = .04 (see Table 2), and indicated there were no significant associations between connectedness and comfort intervening with general bullying situations. With regard to the model covariates (see Table 3), staff in urban settings had lower levels of connectedness than rural and suburban staff. Teachers had lower levels of connectedness than ESPs. In general, elementary staff reported higher levels of connection compared with middle school staff, whereas high school staff reported lower levels of connection compared with those in middle school.

When comfort intervening with special populations was the outcome, Model 1 also demonstrated acceptable fit, CFI = .95, TLI = .95, RMSEA = .03, SRMR = .04 (see Table 2); however, results indicated that greater connectedness was associated with greater comfort intervening with bullying among special populations. As illustrated in Table 2, personal connectedness to the school and staff–staff connectedness were significantly associated with comfort intervening ( $p < .05$ ). The student–staff connectedness was also positively associated with comfort intervening ( $p < .05$ ). Principal connectedness, however, was not significantly associated with comfort intervening with special populations. The associations between covariates and connectedness variables were similar to those in the comfort intervening with general population model (see Table 3).

**Bullying policies and programming.** To examine our second aim, we added the following three school characteristics to the previous models: (a) presence of a district bullying prevention policy, (b) presence of formal bullying prevention programming at the school, and (c) available resources regarding bullying. For comfort intervening with general bullying, the three school characteristics were included (see Model 2 in Table 2) and demonstrated adequate fit (CFI = .93, TLI = .92, RMSEA = .03, SRMR = .08). Results indicated that the availability of resources was associated with feeling more comfortable intervening with general bullying.

The final model (Model 3) examined staff personal involvement and perceptions of bullying resources and programming as they

Table 1  
*Standardized Parameter Estimates for Staff Connectedness and Comfort Intervening Latent Variables*

Latent variable	Estimate
Staff connectedness	
Personal connectedness	
Like to work at school	.75
Ideas listened to	.72
Someone to count on	.53
People care about me	.72
Feel wanted and needed	.84
Feel safe	.55
Recognition for good job	.74
Inspired to do my best	.81
Staff connectedness	
Staff like each other	.85
Staff are friendly with each other	.82
Staff trust and have confidence in each other	.81
Staff help each other	.79
Staff respect each other	.86
Principal connectedness	
Principal shows appreciation	.88
Principal conveys what's expected of staff	.72
Principal looks out for staff	.91
Principal is friendly and approachable	.82
Student connectedness	
Students feel staff are "on their side"	.62
Staff feel pride in the school and its students	.81
Staff really care about students	.68
High expectations for students to achieve	.67
Comfort intervening with special populations	
Sexual orientation or gender nonconformity	.79
Disability	.87
Overweight	.87
Sexist comments	.87
Racist comments	.92
Negative comments about religion	.88
Comfort intervening with general bullying	
Physical	.76
Verbal	.92
Relational	.88
Cyber	.65

Table 2

*Standardized Estimates for Latent Variables in Model and Model Fit Statistics for Special Populations and General Bullying*

Latent variables	Model 1		Model 2		Model 3	
	Special populations	General bullying	Special populations	General bullying	Special populations	General bullying
Personal connectedness	0.05*	0.04	0.05*	0.05*	0.05*	0.07*
Staff connectedness	0.06*	0.04	0.06*	0.06*	0.06*	0.08*
Principal connectedness	-0.002	0.01	0.001	0.02	0.001	0.02
Student connectedness	0.05*	0.02	0.05*	0.03	0.05*	0.10*
District has a policy			0.02	0.03	—	—
Formal prevention efforts			-0.01	0.02	—	—
Resources available			0.12***	0.06***	0.06*	0.10**
Policy easy to implement					0.03	0.04
Received training on policy					0.05*	0.01
Involved in prevention efforts					0.13***	0.13***
Model fit statistics						
Comparative fit index	.95	.94	.94	.93	.94	.93
Tucker-Lewis index	.95	.94	.93	.92	.93	.92
Root-mean-square error of approximation	.026	.028	.028	.031	.028	.027
Standardized root-mean-square residual	.035	.044	.075	.078	.075	.073

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

relate to comfort intervening in general bullying situations. As seen in Table 2, Model 3 demonstrated adequate fit ( $CFI = .93$ ,  $TLI = .92$ ,  $RMSEA = .03$ ,  $SRMR = .07$ ) and indicated that having resources available regarding bullying and being involved in bullying prevention efforts were significantly associated with comfort intervening. However perceiving a school's bullying policy as being easy to implement and receiving training on the schools bullying policy were not significant (see Table 2).

Similar to intervening in general bullying situations, the results of Model 2 indicated that availability of resources was also associated with feeling more comfortable intervening in bullying among special populations. Yet, neither the presence of a district policy nor formal prevention programming in the school was significantly associated with staff's comfort. As seen in Table 2, this model demonstrated adequate fit ( $CFI = .94$ ,  $TLI = .93$ ,  $RMSEA = .03$ ,  $SRMR = .04$ ). The final model (Model 3) examined staff personal involvement and perceptions of bullying re-

sources and programming in combination with the retained variables from Models 1 and 2. Model 3 demonstrated similar fit to the prior model ( $CFI = .94$ ,  $TLI = .93$ ,  $RMSEA = .03$ ,  $SRMR = .07$ ). Results indicated that having resources available regarding bullying, receiving training on the school's bullying policy, and being involved in bullying prevention efforts were significantly associated with comfort intervening; however, perceiving the policy as being easy to implement was not (see Table 2).

**Model covariates.** Several covariates were included in the general bullying and special populations models to control for staff characteristics as they relate to the latent connectedness variables (see Table 3). For both the general bullying and special populations models, teachers (compared with ESPs), high school staff, and staff working in urban neighborhoods tended to have lower levels of connectedness ( $p < .05$ ). Surprisingly, the amount of interaction between staff and students was not significantly related to their level of connectedness. In terms of personal characteristics,

Table 3

*Standardized Estimates for Covariates Included in Model 1*

Covariates	Personal connectedness		Staff connectedness		Principal connectedness		Student connectedness	
	Special populations	General bullying	Special populations	General bullying	Special populations	General bullying	Special populations	General bullying
Amount of student interaction	0.02	0.02	0.02	0.02	-0.02	-0.01	0.00	0.00
Survey mode	0.18***	0.17***	0.20***	0.16***	0.26***	0.14***	0.15***	0.15***
Urban	-0.16***	-0.14***	-0.15***	-0.12***	-0.15***	-0.08**	-0.18***	-0.17***
Teachers (vs. ESPs)	-0.08***	-0.06***	-0.05*	-0.03*	-0.16***	-0.07***	-0.05***	-0.04*
Elementary school	0.07**	0.07**	0.05	0.05	0.05	0.03	0.18***	0.20***
High school	-0.12***	-0.11***	-0.12***	-0.10***	-0.12*	-0.07*	-0.09**	-0.10**
Years worked	0.001	-0.003	0.001	-0.003	-0.002	-0.008	0.002	-0.001
Age	-0.001	-0.003	0.002	-0.001	-0.001	-0.004	0.002	0.000

Note. Middle school served as the reference group for the elementary and high school covariates. ESPs = education support professionals.

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .



staff who were older and had spent more years working in education tended to have lower levels of connectedness, but this was only the case for the general bullying model ( $p < .01$ ). Finally, we conducted a set of exploratory post hoc interactions for urbanicity and school level (in lieu of modeling them as covariates). We constrained the association between the latent school connectedness variables and the outcome to be equal across groups and examined the difference in model fit via the Wald test (Muthén & Muthén, 1998–2012); however, we found no evidence of a significant interaction effect involving either of these variables (results not reported). Therefore, the exploratory interaction effects were dropped from the final models.

## Discussion

The current challenge in bullying prevention is both reducing rates of violence at school while also bolstering school climate and student and staff members' connection to the school community (Waasdorp, Bradshaw, & Leaf, 2012). Despite the surge in the literature on these two interrelated topics, relatively little empirical research has specifically examined the overlap between bullying prevention and school connectedness from the perspective of school staff. The purpose of the current study was to tease apart the multiple domains of staff connectedness and examine how they directly relate to the comfort of school staff in intervening in bullying situations. In addition, we examined how schoolwide policies and programming influence staff members' likelihood of intervening. This line of research has considerable relevance for educational psychologists interested in improving conditions for learning and engaging school staff in prevention efforts.

### Staff Connectedness and Comfort Intervening

We found that the relationship between staff connectedness and comfort intervening varied depending on whether it was a general bullying situation or one that involved special populations of students. For general bullying, personal, student, and peer connectedness were salient factors in staff members' comfort intervening only when schoolwide policies and programming efforts were added to the model. On the other hand, higher levels of staff connectedness were consistently related to reports of being more comfortable intervening with special populations. Specifically, staff members' close relationships with students and their colleagues, as well as the school in general had a positive impact on their comfort intervening with at-risk groups of students. This finding is consistent with the educational research suggesting that trust and support are the key elements to creating successful working relationships within the school (Skaalvik & Skaalvik, 2011; Wahlstrom & Louis, 2008). These positive working relationships likely help to create a collective sense of school pride that encourages school personnel to take a proactive stance on bullying prevention.

However, not all aspects of connectedness were associated with staff members' comfort intervening. Of the four domains of connectedness, staff relationships with administration were not predictive of comfort intervening in general bullying and special populations. This is somewhat surprising since research shows that when teachers and staff feel supported by their administration, they tend to report higher levels of commitment and more collegiality,

the consequence of which is increased staff retention (Singh & Billingsley, 1998). Perhaps, teachers' relationship with the school's administration team is more salient with regard to schoolwide program buy-in but not necessarily for on-the-spot bullying intervention. Additional research is needed to examine how administrative support impacts staff perceptions of the school and involvement in bullying intervention efforts. In contrast, although few specific hypotheses were formulated related to the staff connectedness with colleagues, this form of connectedness was associated with intervening in general bullying situations and special populations. Taken together, these findings emphasize the importance of multiple domains of connectedness in influencing staff members' willingness to intervene in bullying situations.

**School factors and connectedness.** The analyses also highlighted some salient school-level factors that influence staff reports of connectedness. This is particularly important given prior research suggesting that characteristics of the school environment are linked to teacher willingness to implement programs in their classroom (Domitrovich et al., 2008; Han & Weiss, 2005). For example, staff in urban settings had lower levels of connectedness than schools located in rural and suburban settings. Schools in urban inner-city neighborhoods are typically at a greater economic disadvantage, have less social cohesion, and have fewer resources for educating children than suburban schools (Elliott et al., 1996; Tolan, Gorman-Smith, & Henry, 2003). With the increased risk for schoolwide disorganization, urban schools also tend to have higher rates of disruptive student behavior. A national survey of teachers revealed that teachers working in urban communities report more problem behavior in their classrooms than teachers working in suburban and rural schools (Provasnik et al., 2007). Thus, it is important for these schools to tailor bullying interventions that both teach staff how to intervene effectively and build connectedness and trust among the school community.

Second, there were several significant differences by grade level, such that elementary staff reported the highest level of connectedness, followed by middle, and then high school staff members; however, the associations were not moderated by school level. Prior research on student connectedness found similar results, with younger students having higher levels of school connectedness than their older peers (Furlong, Pavelski, & Saxton, 2002; Whitlock, 2006). Elementary teachers tend to develop closer relationships with their students since they stay in one class the majority of the school day, whereas middle and high school-age youth frequently transition from class to class. Middle and high schools also tend to have higher enrollments and larger school campuses, which reduces the ability of staff members to form close, personal relationships with one another. One way that schools can foster connectedness among teaching staff is by involving staff across grade levels in schoolwide programming efforts, as opposed to specific grade-level meetings.

### Bullying Programming and Comfort Intervening

A secondary aim of the study was to examine the link between staff perceptions of their school's bullying prevention programming and their comfort intervening with bullying. It is interesting that the presence of a district bullying policy and its ease of implementation, as well as the availability of formal prevention programming at the school, were not associated with comfort



intervening in bullying. In other words, simply having policies and programs available does not equate to teachers and staff feeling comfortable implementing programs with fidelity. Rather, the analyses show that staff who are actively engaged with schoolwide programming are more likely to help those youth involved in bullying. Specifically, when staff were involved in the bullying prevention efforts and had access to bullying resources at their school, they tended to feel more comfortable intervening in both general forms of bullying, as well as bullying related to sexual orientation or gender nonconformity, disability, being overweight, sexism, racism, and religion. Previous research indicates that bullying prevention programs are not only more effective but are more likely to be sustained over time if staff and administrators take part in developing the program (Hirschstein & Frey, 2006; Rigby, 2007). In a longitudinal study, Rhodes, Camic, Milburn, and Lowe (2009) found that when school staff were active collaborators in identifying schoolwide programs (e.g., antibullying, after-school activities, and teacher wellness programs) and subsequently assisted in the implementation of these interventions, there were increases in teacher attitudes and perceptions of school climate. More important, these changes in perceptions of climate were found to positively impact students' perceptions of the school (e.g., positive peer interactions, teacher support, and academic engagement). Therefore, prior to implementing a prevention program, it would be important to assess the level of engagement and interest from teachers and school staff in order to ensure program effectiveness.

Drawing upon the education literature, studies have shown teachers' level of experience and prior training influence their perceptions of self-efficacy to teach students (e.g., Tschannen-Moran & Woolfolk Hoy, 2007). This line of research suggests that when teachers feel that they have the skills to intervene in bullying or perceive they can help change the school norms related to peer victimization (Wood, 1992; Domitrovich et al., 2008), the more likely it is that they will take a proactive stance on reducing bullying at school. A growing body of research has shown that when teachers and administrators work collaboratively on school-based prevention programs, teachers tend to be more invested in the program's short- and long-term outcomes (Adelman & Taylor, 2003). Taken together, these findings suggest that both students' and staff members' perceptions of school climate may be predictive of bullying prevention program implementation and outcomes (Beets et al., 2008; Bradshaw et al., 2009; Bradshaw & Waasdorp, 2009).

### Limitations and Strengths

It is important to note some limitations when interpreting these findings. For example, the data are self-reported; therefore, we are unable to ensure the validity of these data. Future research should employ a mixed-methods approach, combining teacher report and observational data to capture how connectedness relates to staff responses to bullying incidents. Social desirability may play a role in participants' responses, and this may vary by the mode of survey (i.e., Web vs. phone). Not all participants e-mailed or called agreed to participate; therefore, it is possible that staff more involved in bullying prevention efforts or more concerned about the issue were more likely to agree to participate. The data are cross-sectional, so we are unable to draw any conclusions regarding

causality. Additional research is needed to examine other factors, such as teacher efficacy and burnout on teachers' willingness to intervene.

Nevertheless, this study has several strengths, most notably the nationally representative design, the large sample size, the linkage with the NEA population, the use of propensity scores to address potential sampling biases, and the inclusion of teaching and non-teaching school staff. Due to the sampling strategy employed, the respondents were not nested within schools, and there was very little nesting within districts; therefore, multilevel analysis was not warranted (Raudenbush & Bryk, 2002). Additional school-level factors and student perspectives not assessed in the current study (e.g., school climate) could provide a more comprehensive view of staff perceptions and responses to bullying and possibly allow for multilevel analyses (O'Brennan, Bradshaw, & Furlong, in press). There may also be differences in perceptions among the subpopulations of ESPs (e.g., transportation workers and paraprofessionals), which will be investigated in future studies. Although we adjusted for region and several school characteristics, it is possible that district- or state-level factors (e.g., policies and laws regarding bullying) not examined in this study may have influenced the pattern of findings. This is another area for investigation in future studies.

### Conclusions and Implications for Educational Research

These findings also highlight the importance of connectedness particularly in relation to intervening in bullying situations involving special populations. Specifically, the findings suggest that the same factors associated with intervening in general bullying situations were relevant for bullying involving special populations, with the exception of receiving training on bullying policies, which was relevant to bullying among special populations but not for bullying in general. This suggests that bullying prevention policies may signal particular relevance for bullying targeting particularly sensitive populations.

The results also suggest that connectedness may be an important target for bullying prevention programming and climate promoting efforts. It is likely that connectedness, specifically connection to students, colleagues, and the larger school community, increases staff members' willingness to intervene in bullying situations. Connectedness-promoting activities may enhance staff's dedication toward making the school community a positive and safe atmosphere for their colleagues and students and, in turn, increase their empathy for youth involved in bullying. Connectedness efforts also have the potential of increasing job satisfaction and retention, which is a major concern given the high rate of turnover in the field of education (Boe, Cook, & Sunderland, 2008). Recent national data suggest that 10% of public school teachers leave the profession after 1 year, and an additional 12% leave after 2 years of teaching (Kaiser, 2011). Moreover, teacher attrition hinders the overall social milieu of the school and limits the fidelity of program implementation from year-to-year (Borman & Dowling, 2008; National Commission on Teaching and America's Future, 2007). However, if schools are able to foster support and trust among staff members, they are more likely to reduce rates of bullying and implement programs with efficacy. Although not examined in the current study, it is likely that these associations



generalize to other staff activities, such as efforts to improve conditions for learning as well as quality implementation of prevention programs (Domitrovich et al., 2008).

It is important to remember, however, that the school climate improvement process is slow and likely requires a change in norms and behavior (Bradshaw et al., 2009). Identifying factors associated with positive schoolwide climate and behavioral changes can inform the development of programs and policies related to school safety and bullying prevention. Therefore, identifying school contextual factors associated with behavior and school climate change would greatly inform the bullying prevention literature. Contextual factors, such as the school's organizational climate or the level of disorder within the school or classroom environment, may also influence the way in which teachers manage bullying and other discipline problems, participate in whole-school prevention efforts, or refer students to school-based services (e.g., counseling; Domitrovich et al., 2008). Further work is needed to determine efficient ways to assess readiness and help schools move toward fidelity and positive student outcome, and toward strategies to program sustainability.

## References

- Adelman, H. S., & Taylor, L. (2003). On sustainability of project innovations as systemic change. *Journal of Educational & Psychological Consultation, 14*, 1–25. doi:10.1207/S1532768XJEPC1401\_01
- Beets, M. W., Flay, B. R., Vuckinich, S., Acock, A. C., Li, K., & Allred, C. (2008). School climate and teachers' beliefs and attitudes associated with implementation of the Positive Action Program: A diffusion of innovations model. *Prevention Science, 9*, 264–275. doi:10.1007/s11121-008-0100-2
- Bevans, K. B., Bradshaw, C. P., Miech, R., & Leaf, P. J. (2007). Staff- and school-level predictors of school organizational health: A multilevel analysis. *Journal of School Health, 77*, 294–302. doi:10.1111/j.1746-1561.2007.00210.x
- Boe, E. E., Cook, L. H., & Sunderland, R. J. (2008). Teacher turnover: Examining exit attrition, teaching area transfer, and school migration. *Exceptional Children, 75*, 7–31.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley Interscience.
- Borman, G. D., & Dowling, N. M. (2008). Attrition and retention: A meta-analytic and narrative review of the research. *Review of Educational Research, 78*, 367–409. doi:10.3102/0034654308321455
- Bradshaw, C. P., Koth, C. W., Thornton, L. A., & Leaf, P. J. (2009). Altering school climate through school-wide Positive Behavioral Interventions and Supports: Findings from a group-randomized effectiveness trial. *Prevention Science, 10*, 100–115. doi:10.1007/s11121-008-0114-9
- Bradshaw, C. P., Reinke, W. M., Brown, L. D., Bevans, K. B., & Leaf, P. J. (2008). Implementation of school-wide Positive Behavioral Interventions and Supports (PBIS) in elementary schools: Observations from a randomized trial. *Education & Treatment of Children, 31*, 1–26. doi:10.1353/etc.0.0025
- Bradshaw, C. P., Sawyer, A. L., & O'Brennan, L. M. (2007). Bullying and peer victimization at school: Perceptual differences between students and school staff. *School Psychology Review, 36*, 361–382.
- Bradshaw, C. P., & Waasdorp, T. E. (2009). Measuring and changing a "culture of bullying." *School Psychology Review, 38*, 356–361.
- Bradshaw, C. P., Waasdorp, T. E., & O'Brennan, L. (2010). *NEA members' knowledge and experience with bullying questionnaire*. Washington, DC: National Education Association.
- Bradshaw, C. P., Waasdorp, T. E., O'Brennan, L., & Gulemetova, M. (2013). Teachers' and education support professionals' perspectives on bullying and prevention: Findings from a National Education Association (NEA) survey. *School Psychology Review, 42*, 280–297.
- Brixval, C. S., Rayce, S. L. B., Rasmussen, M., Holstein, B. E., & Due, P. (2012). Overweight, body image and bullying: An epidemiological study of 11- to 15-years olds. *European Journal of Public Health, 22*, 126–130. doi:10.1093/eurpub/ckr010
- Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Cambridge, MA: Harvard University of Press.
- Bronfenbrenner, U., & Morris, P. (1998). The ecology of developmental processes. In W. Damon (Ed.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (pp. 993–1028). New York, NY: Wiley.
- Brown, K. E., & Medway, F. J. (2007). School climate and teacher beliefs in a school effectively serving poor South Carolina African-American students: A case study. *Teaching and Teacher Education, 23*, 529–540. doi:10.1016/j.tate.2006.11.002
- Butler, R. (2012). Striving to connect: Extending an achievement goal approach to teacher motivation to include relational goals for teaching. *Journal of Educational Psychology, 104*, 726–742. doi:10.1037/a0028613
- Collie, R. J., Shapka, J. D., & Perry, N. E. (2011). Predicting teacher commitment: The impact of school climate and social-emotional learning. *Psychology in the Schools, 48*, 1034–1048. doi:10.1002/pits.20611
- Collie, R. J., Shapka, J. D., & Perry, N. E. (2012). School climate and social-emotional learning: Predicting teacher stress, job satisfaction, and teaching efficacy. *Journal of Educational Psychology, 104*, 1189–1204. doi:10.1037/a0029356
- Doll, B., Song, S., & Siemers, E. (2004). Classroom ecologies that support or discourage bullying. In D. L. Espelage & S. M. Swearer (Eds.), *Bullying in American schools: A social-ecological perspective on prevention and intervention* (pp. 161–183). Mahwah, NJ: Erlbaum.
- Domitrovich, C. E., Bradshaw, C. P., Poduska, J., Hoagwood, K., Buckley, J., Olin, S., . . . Ialongo, N. S. (2008). Maximizing the implementation quality of evidence-based preventive interventions in schools: A conceptual framework. *Advances in School Mental Health Promotion: Training and Practice, Research and Policy, 1*, 6–28.
- Elliott, D. S., Wilson, W. J., Huizinga, D., Sampson, R. J., Elliott, A., & Rankin, B. (1996). The effects of neighborhood disadvantage on adolescent development. *Journal of Research in Crime and Delinquency, 33*, 389–426. doi:10.1177/0022427896033004002
- Furlong, M. J., Pavelski, R. E., & Saxton, J. D. (2002). The prevention of school violence. In S. E. Brock, P. J. Lazarus, & S. R. Jimerson (Eds.), *Crisis response in the schools* (pp. 131–139). Washington, DC: National Association of School Psychologists.
- Gilman, R., Huebner, E. S., & Furlong, M. J. (2009). *Handbook of positive psychology in schools*. New York, NY: Routledge.
- Greenberg, M. T., Weissberg, R. P., O'Brien, M. U., Zins, J. E., Fredericks, L., Resnik, H., & Elias, M. J. (2003). Enhancing school-based prevention and youth development through coordinated social, emotional, and academic learning. *American Psychologist, 58*, 466–474. doi:10.1037/0003-066X.58.6-7.466
- Gregory, A., Henry, D. B., Schoeny, M. E., & the Metropolitan Area Child Study Research Group. (2007). School climate and implementation of a preventive intervention. *American Journal of Community Psychology, 40*, 250–260. doi:10.1007/s10464-007-9142-z
- Han, S. S., & Weiss, B. (2005). Sustainability of teacher implementation of school-based mental health programs. *Journal of Abnormal Child Psychology, 33*, 665–679. doi:10.1007/s10802-005-7646-2
- Hirschstein, M. K., & Frey, K. S. (2006). Promoting behavior and beliefs that reduce bullying: The Steps to Respect program. In S. Jimerson & M. Furlong (Eds.), *The handbook of school violence and school safety: From research to practice* (pp. 309–324). Mahwah, NJ: Erlbaum.

- Holbrook, A. L., Krosnick, J. A., & Pfent, A. (2007). The causes and consequences of response rates in surveys by the news media and government contractor survey research firms. In M. Lepkowski, C. Tucker, J. M. Brick, E. D. de Leeuw, L. Japac, P. J. Lavrakas, . . . R. L. Sangster (Eds.), *Advances in telephone survey methodology* (pp. 499–678). doi:10.1002/9780470173404.ch23
- Hoy, W. K., & Woolfolk, A. E. (1993). Teachers' sense of efficacy and the organizational health of schools. *Elementary School Journal*, 93, 355–372. doi:10.1086/461729
- Johnson, W. L., Johnson, A. M., Kranch, D. A., & Zimmerman, K. J. (1999). The development of a university version of the Charles F. Kettering Climate Scale. *Educational and Psychological Measurement*, 59, 336–350. doi:10.1177/00131649921969884
- Kaiser, A. (2011). *Beginning teacher attrition and mobility: Results from the first through third waves of the 2007–08 Beginning Teacher Longitudinal Study* (NCES 2011–318). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved August 2, 2012, from <http://nces.ed.gov/pubsearch>
- Kallestad, J. H., & Olweus, D. (2003). Predicting teachers' and schools' implementation of the Olweus Bullying Prevention Program: A multi-level study. *Prevention and Treatment*, 6, Article 21.
- Keigher, A. (2009). *Characteristics of public, private, and bureau of Indian education elementary and secondary schools in the United States: Results From the 2007–08 Schools and Staffing Survey* (NCES 2009–321). Washington, DC: U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.
- Kochenderfer-Ladd, B., & Pelletier, M. E. (2008). Teachers' views and beliefs about bullying: Influences on classroom management strategies and students' coping with peer victimization. *Journal of School Psychology*, 46, 431–453. doi:10.1016/j.jsp.2007.07.005
- Konishi, C., Hymel, S., Zumbo, B. D., & Li, Z. (2010). Do school bullying and student-teacher relationships matter for academic achievement? A multilevel analysis. *Canadian Journal of School Psychology*, 25, 19–39. doi:10.1177/0829573509357550
- Kosciw, J. G., Greytak, E. A., Diaz, E. M., & Bartkiewicz, M. J. (2010). *The 2009 National School Climate Survey: The experiences of lesbian, gay, bisexual and transgender youth in our nation's schools*. New York, NY: Gay Lesbian, and Straight Education Network.
- Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and Web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, 72, 847–865. doi:10.1093/poq/nfn063
- Libbey, H. P. (2004). Measuring student relationships to school: Attachment, bonding, connectedness, and engagement. *Journal of School Health*, 74, 274–283. doi:10.1111/j.1746-1561.2004.tb08284.x
- Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- Nansel, T. R., Overpeck, M., Pilla, R. S., Ruan, W. J., Simons-Morton, B., & Scheidt, P. (2001). Bullying behaviors among U.S. youth: Prevalence and associations with psychosocial adjustment. *Journal of the American Medical Association*, 285, 2094–2100. doi:10.1001/jama.285.16.2094
- National Commission on Teaching and America's Future. (2007). *Policy brief: The high cost of teacher turnover*. Washington, DC: Author.
- O'Brennan, L. M., Bradshaw, C. P., & Furlong, M. J. (in press). Influence of classroom and school climate on teacher perceptions of student problem behavior. *School Mental Health*.
- O'Brennan, L. M., & Furlong, M. F. (2010). Relations between students' perceptions of school connectedness and peer victimization. *Journal of School Violence*, 9, 375–391. doi:10.1080/15388220.2010.509009
- Olweus, D. (1993). Bully/victim problems among schoolchildren: Long-term consequences and an effective intervention program. In S. Hodgins (Ed.), *Mental disorder and crime* (pp. 317–349). Newbury Park, CA: Sage.
- Olweus, D., Limber, S., & Mihalic, S. F. (1999). *Blueprints for Violence Prevention Series: Book 9. Bullying prevention program*. Boulder: University of Colorado, Institute of Behavioral Science, Center for the Study and Prevention of Violence.
- Parker, P. D., Martin, A. J., Colmar, S., & Liem, G. A. (2012). Teachers' workplace well-being: Exploring a process model of goal orientation, coping behavior, engagement, and burnout. *Teaching and Teacher Education*, 28, 503–513. doi:10.1016/j.tate.2012.01.001
- Pas, E. T., Bradshaw, C. P., Hershfeldt, P. A., & Leaf, P. J. (2010). A multilevel exploration of the influence of teacher efficacy and burnout on response to student problem behavior and school-based service use. *School Psychology Quarterly*, 25, 13–27. doi:10.1037/a0018576
- Provasnik, S., KewalRamani, A., Coleman, M. M., Gilbertson, L., Herring, W., & Xie, Q. (2007). *Status of education in rural America* (NCES 2007–040). Washington, DC: National Center for Education Statistics.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Rhodes, J. E., Camic, P. M., Milburn, M., & Lowe, S. R. (2009). Improving middle school climate through teacher-centered change. *Journal of Community Psychology*, 37, 711–724. doi:10.1002/jcop.20326
- Richard, J. F., Schneider, B. H., & Mallet, P. (2012). Revisiting the whole-school approach to bullying: Really looking at the whole school. *School Psychology International*, 33, 263–284. doi:10.1177/0143034311415906
- Rigby, K. (2007). *Children and bullying: How parents and educators can reduce bullying at school*. Malden, MA: Wiley-Blackwell.
- Rose, C. A., Monda-Amaya, L. E., & Espelage, D. L. (2011). Bullying perpetration and victimization in special education: A review of the literature. *Remedial and Special Education*, 32, 114–130. doi:10.1177/0741932510361247
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55. doi:10.1093/biomet/70.1.41
- Russell, S. T., Ryan, C., Toomey, R. B., Diaz, R. M., & Sanchez, J. (2011). Lesbian, gay, bisexual, and transgender adolescent school victimization: Implications for young adult health and adjustment. *Journal of School Health*, 81, 223–230. doi:10.1111/j.1746-1561.2011.00583.x
- Sawyer, A. L., Bradshaw, C. P., & O'Brennan, L. M. (2008). Examining ethnic, gender, and developmental differences in the way children report being a victim of "bullying" on self-report measures. *Journal of Adolescent Health*, 43, 106–114. doi:10.1016/j.jadohealth.2007.12.011
- Schonlau, M., van Soest, A., Kapteyn, A., & Couper, M. (2009). Selection bias in web surveys and the use of propensity scores. *Sociological Methods & Research*, 37, 291–318. doi:10.1177/0049124108327128
- Singh, K., & Billingsley, B. (1998). Professional support and its effects on teachers' commitment. *Journal of Educational Research*, 91, 229–239. doi:10.1080/00220679809597548
- Skaalvik, E. M., & Skaalvik, S. (2011). Teacher job satisfaction and motivation to leave the teaching profession: Relations with school context, feeling of belonging, and emotional exhaustion. *Teaching and Teacher Education*, 27, 1029–1038. doi:10.1016/j.tate.2011.04.001
- Sun, R. C. F., Shek, D. T. L., & Siu, A. M. H. (2008). Positive school and classroom environment: Precursors of successful implementation of positive youth development programs. *Scientific World Journal*, 8, 1063–1074. doi:10.1100/tsw.2008.126
- Swearer, S. M., Espelage, D. L., Vaillancourt, T., & Hymel, S. (2010). What can be done about school bullying? Linking research to educational practice. *Educational Researcher*, 39, 38–47. doi:10.3102/0013189X09357622



- Taylor, H. (2000). Does Internet research "work"? Comparing on-line survey results with telephone surveys. *International Journal of Market Research*, 42, 51–63.
- Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A review of school climate research. *Review of Educational Research*, 83, 357–385. doi:10.3102/0034654313483907
- Tolan, P. H., Gorman-Smith, D., & Henry, D. B. (2003). The developmental ecology of urban males' youth violence. *Developmental Psychology*, 39, 274–291. doi:10.1037/0012-1649.39.2.274
- Tschannen-Moran, M., & Woolfolk Hoy, A. (2007). The differential antecedents of self-efficacy beliefs of novice and experienced teachers. *Teaching and Teacher Education*, 23, 944–956. doi:10.1016/j.tate.2006.05.003
- Waasdorp, T. E., Bradshaw, C. P., & Leaf, P. J. (2012). The impact of Schoolwide Positive Behavioral Interventions and Supports (SWPBIS) on bullying and peer rejection: A randomized controlled effectiveness trial. *Archives of Pediatrics and Adolescent Medicine*, 166, 149–156. doi:10.1001/archpediatrics.2011.755
- Wahlstrom, K. L., & Louis, K. S. (2008). How teachers experience principal leadership: The roles of professional community, trust, efficacy, and shared responsibility. *Educational Administration Quarterly*, 44, 458–495. doi:10.1177/0013161X08321502
- Watts, M. (2010). *An experiment and analysis in Web, phone, and multi-modal surveying of NEA members*. Northampton, MA: Abacus.
- Whitlock, J. L. (2006). Youth perceptions of life at school: Contextual correlates of school connectedness in adolescence. *Applied Developmental Science*, 10, 13–29. doi:10.1207/s1532480xads1001\_2
- Wood, C. J. (1992). Effect of teacher involvement and teacher self-efficacy on ratings of consultant effectiveness and intervention acceptability. *Journal of Educational & Psychological Consultation*, 3, 301–316. doi:10.1207/s1532768xjepc0304\_2
- You, S., O'Malley, M., & Furlong, M. (2013). Preliminary development of the Brief-California School Climate Survey: Dimensionality and measurement invariance across teachers and administrators. *School Effectiveness and School Improvement*, 24, 1–21. doi:10.1080/09243453.2013.784199
- Zablotsky, B., Bradshaw, C. P., Anderson, C., & Law, P. (2012). Involvement in bullying among children with autism spectrum disorders: Parents' perspectives on the influence of school factors. *Behavioral Disorders*, 37, 179–191.
- Zullig, K. J., Koopman, T. M., Patton, J. M., & Ubbes, V. A. (2010). School climate: Historical review, instrument development, and school assessment. *Journal of Psychoeducational Assessment*, 28, 139–152. doi:10.1177/0734282909344205

Received April 29, 2013

Revision received December 10, 2013

Accepted December 29, 2013 ■

# Testing the Theory of Successful Intelligence in Teaching Grade 4 Language Arts, Mathematics, and Science

Robert J. Sternberg  
Cornell University

Linda Jarvin  
Paris College of Art

Damian P. Birney  
University of Sydney

Adam Naples  
Yale University

Steven E. Stemler  
Wesleyan University

Tina Newman  
Center for Children With Special Needs,  
Glastonbury, Connecticut

Renate Otterbach  
University of San Francisco

Carolyn Parish  
SRA International, Fairfax, Virginia

Judy Randi  
University of New Haven

Elena L. Grigorenko  
Yale University

This study addressed whether prior successes with educational interventions grounded in the theory of successful intelligence could be replicated on a larger scale as the primary basis for instruction in language arts, mathematics, and science. A total of 7,702 4th-grade students in the United States, drawn from 223 elementary school classrooms in 113 schools in 35 towns (14 school districts) located in 9 states, participated in the program. Students were assigned, by classroom, to receive units of instruction that were based either upon the theory of successful intelligence (SI; analytical, creative, and practical instruction) or upon teaching as usual (weak control), memory instruction (strong control), or critical-thinking instruction (strong control). The amount of instruction was the same across groups. In the 23 comparisons across 10 content units in 3 academic domains, there were only a small number of instances in which students in the SI instructional groups generally performed statistically better than students in other conditions. There were even fewer instances where the different control conditions outperformed the SI students. Implications for the future of SI theory and the scalability of research efforts in general are discussed.

*Keywords:* successful intelligence, critical thinking, memory, instruction, scalability

Throughout the first decade of the 21st century, educational researchers and policymakers have placed an increased emphasis on the twin goals of (a) using experimental designs to evaluate educational interventions and (b) gaining a greater understanding of the issues related to the scalability of educational interventions. The value

placed on interventions that have been experimentally tested is highlighted by repositories such as the U.S. Department of Education's "What Works" clearinghouse (<http://ies.ed.gov/ncee/wwc/>). Projects related to issues of scalability were funded by the Department of Education in the early to mid-2000s, and the results of these projects

---

This article was published Online First April 7, 2014.

Robert J. Sternberg, Department of Human Development, Cornell University; Linda Jarvin, Paris College of Art; Damian P. Birney, School of Psychology, University of Sydney; Adam Naples, Child Development Center, Yale University; Steven E. Stemler, Department of Psychology, Wesleyan University; Tina Newman, Center for Children With Special Needs, Glastonbury, Connecticut; Renate Otterbach, Department of General Education, University of San Francisco; Carolyn Parish, SRA International, Fairfax, Virginia; Judy Randi, Department of Education, University of New Haven; Elena L. Grigorenko, Child Study Center and Department of Psychology, Yale University.

This research was supported primarily by National Science Foundation Grant REC-9979843 with additional support from the Javits Act Program (Grant No. R206R000001). We are grateful to Sig Abeles, Jill Citron-Pousty, William Disch, Tona Donlon, Sarah Duman, Rebecca Felton, PJ Henry, Alex Isgut, Steve Leinwand, Delci Lev, Donna Macomber, Mari Muri, Nefeli Misuraca, Paul O'Keefe, Alina Reznitskaya, Robyn Rissman, Morgan Reynolds, Christina Schwarz, Emma Seppala, Gregory Snorheim, Heidi Soxman, Cheri Stahl, and Olga Stepanosova for their invaluable assistance on this project.

Correspondence concerning this article should be addressed to Robert J. Sternberg, Department of Human Development, Cornell University, B44 MVR Hall, Ithaca, NY 14853. E-mail: [robert.sternberg@cornell.edu](mailto:robert.sternberg@cornell.edu)



are beginning to receive increased attention in the empirical literature (Constas & Sternberg, 2006; McKenna & Walpole, 2010).

In the present article, we report on a large-scale empirical field study that also sought to address issues related to scalability. We examined whether applying the theory of successful intelligence to instruction and assessment in Grade 4 language arts, mathematics, and science would result in superior learning outcomes relative to alternative instructional methods, in particular, memory-based instruction and critical-thinking based instruction (strong comparison/control conditions) and teaching as usual—whatever it happened to be (weak comparison/control condition). The study involved the participation of 7,702 fourth-grade students in 113 elementary schools and 223 classrooms across the United States in 35 towns (14 school districts) located in nine states (Alabama, California, Connecticut, Massachusetts, Minnesota, Kansas, North Carolina, South Carolina, and West Virginia), in order to determine whether prior successes with the theory's instructional application could be replicated at scale.

## Background

There is evidence to suggest that teaching and assessment may be more effective when they are based in part on cognitive-psychological theories that have been applied to education (Bruning, Schraw, & Norby, 2010; Corno, Cronbach, Kupermintz, & Lohman, 2001). Certainly, this has been a major claim of researchers as well as textbook authors in educational psychology (e.g., Ormrod, 2010; Slavin, 2008; Woolfolk, 2009). One such cognitive-psychological theory is the theory of successful intelligence.

The theory (Sternberg, 1997, 2005, 2010) argues that *successful intelligence* is a person's ability to achieve his or her goals in life, within his or her sociocultural context, by capitalizing on strengths and correcting or compensating for weaknesses, in order to adapt to, shape, and select environments through a combination of analytical, creative, and practical skills (Sternberg, 2003b, 2009; Sternberg, Grigorenko, & Jarvin, 2007; Sternberg, Jarvin, & Grigorenko, 2009). Different students have different combinations of these skills. The theory is based on the notion that students learn in different ways and that they have different strengths in learning (Sternberg, Grigorenko, & Zhang, 2008a, 2008b), just as teachers have different strengths in teaching (Spear & Sternberg, 1987). Our goal is to assist teachers in balancing their teaching in such a way that each of the abilities can be addressed, exercised, and given a chance to develop (Sternberg & Grigorenko, 2000; Sternberg et al., 2007, 2009).

Teaching for analytical thinking means encouraging students to (a) analyze, (b) critique, (c) judge, (d) compare and contrast, (e) evaluate, or (f) assess. When teachers refer to teaching for "critical thinking," some of them may mean teaching for analytical thinking. Examples of exercises designed to develop such skills might ask students to (a) analyze a political speech, (b) critique a work of art, (c) judge the value of a social program, (d) compare and contrast two works of literature, (e) evaluate the conclusions drawn from a scientific experiment, or (f) assess the rationale for a cultural custom.

Teaching for creative thinking means encouraging students to (a) create, (b) invent, (c) discover, (d) imagine if . . . , (e) suppose that . . . , (f) predict . . . , or (g) design. Teaching for creative

thinking requires teachers not only to support and encourage creativity but also to role-model it and to reward it when it is displayed (Sternberg & Lubart, 1995; Sternberg & Williams, 1996). Examples of such teaching activities might ask students to (a) create a work of art, (b) invent an alternative ending for a story they read, (c) discover the principle behind a natural phenomenon, (d) imagine what life would be like if global warming continued unabated, (e) suppose that they grew up alingual—having no language at all, (f) predict what will happen in the current civil war in Syria, or (g) design a psychological experiment to test a hypothesis about human behavior.

Teaching for practical thinking means encouraging students to (a) apply, (b) use, (c) put into practice, (d) implement, (e) employ, or (f) persuade someone of something. Such teaching must relate to the real practical needs of the students, not what would be practical for individuals other than the students (Sternberg et al., 2000). Examples might include asking students to (a) apply what they have read in a story to their life, (b) use their knowledge of mathematics to balance a checkbook, (c) put theory into practice in exercising defensive driving, (d) implement a plan for losing (or gaining) weight, (e) employ the rules of haiku and write one, (f) or persuade someone that an argument is sound.

## Measurement Research Support for the Theory of Successful Intelligence

A number of different studies have been conducted that validate the premise of the theory of successful intelligence in the field of assessment and measurement. Here we present them only selectively and briefly.

First, assessments based on the theory of successful intelligence appear to map onto skills that are relevant, broadly speaking, to success in life and various indicators of well-being (e.g., Grigorenko & Sternberg, 2001; Sternberg et al., 2000). Second, these assessments have demonstrated adequate psychometric properties (e.g., Kornilov, Tan, Elliott, Sternberg, & Grigorenko, 2012; Sternberg, Castejón, Prieto, Hautamäki, & Grigorenko, 2001). Third, measurements of different kinds of skills (analytical, creative, and practical) can be done relatively independently of each other (e.g., Grigorenko et al., 2009). Fourth, successful-intelligence assessments can improve prediction of grade-point average as well as prediction of success in extracurricular and leadership activities; such assessments also can reduce ethnic-group differences in performance (Sternberg, 2010; Sternberg, Bonney, Gabora, Karelitz, & Coffin, 2010; Sternberg & The Rainbow Project Collaborators, 2006). Finally, as illustrated in Advanced Placement Psychology, Statistics, and Physics tests, the inclusion of creative and practical assessments in addition to memory and analytical ones can reduce ethnic-group differences while increasing construct validity (Stemler, Grigorenko, Jarvin, & Sternberg, 2006; Stemler, Sternberg, Grigorenko, Jarvin, & Sharpes, 2009).

Thus, there is evidence that assessments based on the theory of successful intelligence can provide valuable concurrent and predictive information about cognitive functioning at various stages of the life span and in various settings.



### Instructional Research Support for the Theory of Successful Intelligence

A number of instructional studies have been conducted with students in different age groups and in various subjects to validate the relevance of the theory of successful intelligence in the classroom (for more detail and other research support for the theory, see Sternberg, 1985, 1997, 2003b; Sternberg, Jarvin, & Grigorenko, 2011). Here we briefly exemplify two types of relevant studies: aptitude–treatment interaction (ATI) and main effect studies of the theory.

An example of the ATI approach is a study (Sternberg, Grigorenko, Ferrari, & Clinkenbeard, 1999) in which the Sternberg Triarchic Abilities Test (STAT; Sternberg, 2003a) was used to assess analytical, creative, and practical skills through multiple-choice and essay items. The test was administered to 326 children around the United States and in some other countries who were identified by their schools as gifted by any standard whatsoever. Children were selected for a summer program in (college-level) psychology if they fell into one of five ability groupings: high analytical, high creative, high practical, high balanced (high in all three abilities), or low balanced (low in all three abilities). The high-school students ( $n = 199$ ) who came to Yale were then divided into four instructional groups. Students in all four instructional groups used the same introductory-psychology textbook, a preliminary version of Sternberg (1995), and listened to the same psychology lectures, by a Yale professor teaching the introduction to psychology course at Yale College. What differed among the four groups was the type of afternoon discussion section to which students were randomly assigned. They were assigned to an instructional condition that emphasized memory, analytical, creative, or practical instruction. The discussion sessions were taught by qualified instructors with no particular training in, or commitment to, the theory of successful intelligence. Instructors were assigned to the instructional conditions at random and were required to use differential teaching approaches. The instructors were unaware of students' patterns of abilities as revealed by the STAT. Consider examples of instruction. In the memory condition, the participants might be asked to recall the originator of a major theory of depression. In the analytical condition, they might be asked to compare and contrast two theories of depression. In the creative condition, they might be asked to formulate their own theory of depression. In the practical condition, they might be asked how they could use what they had learned about depression to help a friend who was depressed. Students in all four instructional conditions were evaluated in terms of their performance on homework, a midterm exam, a final exam, and an independent project. Each type of work was evaluated for analytical, creative, and practical quality. Thus, all students were evaluated in exactly the same way. The results indicated the presence of an aptitude–treatment interaction whereby students who were placed in instructional conditions that better matched their pattern of abilities outperformed students who were mismatched. For all performance assessments combined, for better matched versus mismatched groups, Cohen's  $d$ s were 0.343, 0.195, and 0.255 for analytical, creative, and practical, respectively. In other words, when students are taught at least some of the time in a way that fits how they think, they do better in school. These results suggest that the negative Cronbach and Snow (1977) results for aptitude–treatment

interactions may have been due to lack of theoretical basis for instruction or of theoretical match between instruction and assessment. Pashler, McDaniel, Rohrer, and Bjork (2008), however, have argued that there is still only weak evidence for aptitude–treatment interactions, and the interested reader can refer to Sternberg et al. (2008b) for an alternative point of view.

Subsequently, a main-effect study of the theory (Sternberg, Torff, & Grigorenko, 1998) examined the learning of social studies and science by third graders and eighth graders. The 225 third graders were students in a very low income neighborhood, and the 142 eighth graders were students who were largely middle to upper middle class. Classroom teachers, and consequently their students, were assigned to one of three instructional conditions pseudo-randomly so as to balance the number of students and classrooms in each condition. In the first condition, they were taught the course that they would have learned had there been no intervention (i.e., the emphasis was on memory). In a second condition, teaching emphasized critical (analytical) thinking. In the third condition, students were taught in a way that emphasized a balance of analytical, creative, and practical thinking. All students' performance was assessed for memory learning through multiple-choice assessments as well as for analytical, creative, and practical learning through performance assessments. As expected, students in the successful-intelligence (analytical, creative, practical) condition on average outperformed the other students in terms of the performance assessments. In particular, third graders from the successful-intelligence instructional conditions did better in four out of four comparisons with the standard teaching condition (mean Cohen's  $d = 1.082$  for  $n = 4$ ) and in three out of four comparisons with the critical thinking condition (mean Cohen's  $d = 0.510$  for  $n = 3$ ). Eighth graders in the successful-intelligence condition did better in seven out of seven comparisons with the standard teaching condition (mean Cohen's  $d = 0.842$  for  $n = 7$ ) and in three out of seven comparisons with the critical thinking condition (mean Cohen's  $d = 1.332$  for  $n = 3$ ). One could argue that this result merely reflected the way they were taught. Nevertheless, the result suggested that teaching for these kinds of thinking succeeded. More important, however, was the result that children in the successful-intelligence condition outperformed the other children even on the multiple-choice memory tests (Cohen's  $d$ s were 0.289 and 0.383, and 1.283 and 0.833 for standard and critical thinking instructional conditions in the third- and eighth-grader studies, respectively). In other words, even when the goal is simply to maximize children's memory for information, teaching for successful intelligence is still superior. It enables children to capitalize on their strengths and to correct or to compensate for their weaknesses, allowing them to encode material in a variety of interesting ways.

These results were extended to reading curricula at the middle-school and high-school levels (Grigorenko, Jarvin, & Sternberg, 2002). To illustrate, at the middle-school level ( $n = 871$ ), language arts were taught explicitly for successful intelligence. At the high-school level ( $n = 432$ ), successful intelligence instruction for reading comprehension was infused into instruction in mathematics, physical sciences, social sciences, English, history, foreign languages, and the arts. As in previous studies, each assignment contained analytical, creative, and practical tasks. At both middle- and high-school levels students who were taught for successful intelligence outperformed students who were taught in standard



ways (mean Cohen's  $d = 0.483$  for middle-school level and mean Cohen's  $d = 0.238$  for high-school level).

Ideally, schools might utilize a uniform broad-based, construct-valid, theoretical model in their instruction and assessment and even in admissions, where relevant (Sternberg, 2010). Of course, the model need not be the theory of successful intelligence. Certainly, there are other models (Gardner, 1993, 2006; Mayer, 2011).

The fundamental difference between the current study and the studies discussed above is its scope and specific characteristics. Unlike previous studies, which were framed as either development and narrowly focused efficacy evaluations (Sternberg et al., 1999) or efficacy and replication studies of the theory of successful intelligence in the classroom (Grigorenko et al., 2002; Sternberg et al., 1998), the present study was conceived of as a scaling-up (Sternberg et al., 2006), main-effect evaluation of the utility of the theory of successful intelligence in actual classrooms.

### Scaling up Educational Interventions

Educational research is replete with studies of new and exciting interventions that have been shown to work in one particular context or another. One of the biggest challenges facing the field of educational research, however, is the search for effective interventions (e.g., curricular) that yield similar effects across diverse contexts (Elmore, 1996). In other words, are there interventions that can be successfully scaled up? The concept of "upscaling" is derived from economic theories that are currently pervasive in discussions surrounding education reform in the United States. Specifically, the microeconomic concept of "economies of scale" suggests that certain work can be done more efficiently by increasing the size of operation (Folland, Goodman, & Stano, 2013). In light of such reasoning, several funding agencies, including the National Science Foundation, the Institute for Educational Sciences, and the National Institutes for Health, have been engaged in funding research that has been demonstrated to work in more limited contexts in order to determine whether the results can be replicated on a broader scale (e.g., <http://www.nsf.gov/pubs/2002/nsf02062/nsf02062.pdf>). Their support has funded research by several teams (e.g., Clements, 2005; Francis, 2011; Fuchs, 2004; Hurtig, 2004; Pane, 2007; Starkey, 2004) as well as the present study.

In their book, Glennan, Bodily, Galegher, and Kerr (2004) comprehensively examined the lessons learned from 15 different curricular programs that attempted to go to scale. Generally speaking, the results of this and other research have found three major factors affecting successful scale up (Glennan et al., 2004). First, if the intervention is developed externally, by sources other than teachers themselves, it is often less costly for schools and districts (Nunnery, 1998). This is not to say that teachers have no input. Rather curricula are often co-constructed, with teachers deciding which components to emphasize. Ultimately, however, the easier and less costly it is to implement a design, the more likely it is to be adopted (Glennan et al., 2004). The successful intelligence intervention in the current study was developed externally, although evaluation input from teachers was central to the process (Randi & Jarvin, 2006).

The second factor affecting the success of educational interventions is whether they involve whole-school reform or targeted reform, in which only some classrooms or student populations

receive the intervention. Some evidence suggests that when the whole school is involved, there is greater buy-in across the board, which in turn leads to a greater likelihood of success. The current study represents not only an effort at scaling up but also a large-scale experimental study of different educational interventions. As such, it was neither a targeted reform, per se, nor a traditional whole-school reform.

The third factor impacting the successful scale-up of educational reforms is whether they relate to structural changes, teacher knowledge, or curriculum content. Specifically, prior research has shown that structural changes (e.g., classroom size, student groupings, team teaching) tend to have smaller impacts on educational outcomes than teacher knowledge or curriculum content changes. Within the context of the current study, the focus was on teacher knowledge and curriculum content.

Elias, Zins, Graczyk, and Weissberg (2003) have argued that there is "a need to better document the stories of educational innovation and scaling up efforts so that contextual details can enrich an understanding of what is required for success" (p. 303). The current study is aimed at not only understanding the factors associated with going to scale but also attempting to simultaneously run a large-scale experimental study.

With this work, we attempted to (a) explore whether a curriculum based on the theory of successful intelligence is effective when implemented under conditions that would be typical if a district were to implement it on its own (i.e., without special support from the developer or research team)<sup>1</sup> across a variety of circumstances (e.g., different student populations, different types of schools) and (b) provide an estimate of the robustness of the successful-intelligence instruction. In other words, the main question we sought to address was whether a curriculum based on the theory of successful intelligence would continue to be more effective than instructions relying mostly on memory and/or analytical skills, when implemented on a large scale, with different types of students, school, and teachers. Notably, teacher training and on-going support provided by the research team were much more limited than in previous studies.

## Method

### Participants

Given the scope of this study, our aim was to recruit schools representing a wide range of geographical locations (i.e., different states and different student populations: urban, suburban, and rural), ethnic-minority representation, and socioeconomic profiles. In total, 3,270 school districts across the United States were contacted about the program. The final sample included schools from 35 towns located in 11 counties of nine states (Alabama, California, Connecticut, Massachusetts, Minnesota, Kansas, North Carolina, South Carolina, and West Virginia). We worked in 14 school districts represented by 113 elementary schools, 223 teachers, and 223 classrooms. We entered information on 7,702 student participants, and obtained usable data (i.e., complete pre- and

<sup>1</sup> Of course, only agreeable teachers of classrooms in volunteering schools within the district participated in each condition, and thus the ideal—district implementation—is approximated to varying extents.

posttest data) from 7,574 students. Some students received more than one unit of instruction, but the number of units administered and the order of the units were not fixed and varied depending on the fit between each school district's prescribed content and the topics covered in our units. Correspondingly, here we present the analyses unit by unit, with the total number of observations at  $n = 10,845$ . All students were fourth graders.

Parents and caregivers were informed of the instructional intervention being implemented in their children's classrooms and that the intervention had been endorsed by the school's district superintendent, principal, and classroom teacher. To facilitate broad acceptance, we kept the information collected on individual students to a minimum. Demographic information was thus obtained at the school level. In total, 49.6% of the students in the schools who participated were girls and 27.8% were underrepresented minorities. A further breakdown of the distribution of demographic information across schools by condition is provided in Table 1 (Table 5 provides additional demographic information as it relates to specific units). Study conditions (successful intelligence = SI, critical thinking = CT, memory = M, and teaching as usual = TAU-control) were randomly assigned to schools. Random assignment at the school level was chosen to avoid contamination within a same school building, in which teachers and students naturally talk to each other and share learning materials. In larger districts equal numbers of schools were assigned to each condition, and in small districts, with fewer than three schools participating, the assignment was random.

The guiding principles behind the critical thinking and memory conditions were drawn, respectively, from the education literature (for an introduction, see Halpern, 1996) and research on memory and mnemonic techniques (for an overview, see Baddeley, Eysenck, & Anderson, 2009). As illustrated in Table 2, there is some overlap of activities across conditions, because critical thinking and analytical thinking in the theory of successful intelligence are

similar constructs, and because the corresponding condition included memory activities. The main difference between SI, CT, and M curricula, then, is that the first balances an array of activities whereas the latter two focus on one particular approach (CT or M). Overall, the three versions (SI, CT, and M) have comparable amounts of student activities and require the same duration of classroom time and student time on task to cover the content. Thus, in one case (SI) there is a mixture of different types of activities. In the two other conditions (CT and M) there are more CT and M activities, respectively, and the creative and practical activities that were present in SI are absent.

The study's material development and data collection phase took five years to complete.<sup>2</sup>

## Materials

**Teaching units.** Lesson materials (hereafter, units) were developed for three academic domains (language arts, mathematics, and science) and for different content (e.g., within science there were units on ecology, electricity, light, and magnetism) in a similar manner, equalizing the engagement of targeted skills across the experimental conditions. Each unit was preceded and followed by unit-specific pre- and posttests. The content was based on a thorough review of the standards of each participating state at the time of the creation of the curriculum. We focused on those content topics that a majority of states suggested should be covered in their curriculum at the fourth-grade level. In some cases we selected a topic that was targeted in Grade 4 in one state but in Grade 3 in another state. There was never more than a one-year discrepancy between the topics, however, and, when present, the discrepancy did not influence participation in the study.

The curricula in each of the three instructional treatment conditions (SI, CT, and M) were similar in that they covered the same concepts (e.g., magnetism), contained equal amounts of student activities, and required exact or comparable amounts of classroom instruction and student engagement. They were different, however, in the manner that the concepts were approached, presenting student activities that combined analytical, creative, practical approaches to learning (in the SI condition); or a majority of analytical approaches (in the CT condition); or a majority of activities encouraging memorization (in the M condition). Because the SI instructional approach also engages students in critical thinking, there were some activities that were offered both in the SI curriculum and in the CT curriculum. The same holds true for the memory-based activities that were offered in all three instructional approaches. Table 2 provides an example of how the activities differed in the three instructional approaches: Students in the SI condition had an analytical activity and one practical activity, students in the CT condition had two analytical activities, and

Table 1  
*School-Based Demographic Information Across Intervention Conditions*

Variable	Intervention condition			
	SI	CT	M	TAU-control
% female				
<i>M</i>	49.63	48.21	48.71	47.85
<i>SD</i>	3.71	3.06	2.43	
% Asian				
<i>M</i>	3.83	2.56	3.89	7.38
<i>SD</i>	4.36	3.56	5.12	
% Black				
<i>M</i>	20.39	10.71	28.46	11.38
<i>SD</i>	15.40	14.11	27.01	
% Hispanic				
<i>M</i>	8.97	15.50	9.58	10.77
<i>SD</i>	12.28	29.79	17.11	
% White				
<i>M</i>	66.80	71.23	58.06	70.46
<i>SD</i>	22.24	28.74	29.13	
No. schools	43	40	30	1
No. classes	100	65	55	3

*Note.* SI = successful intelligence; CT = critical thinking; M = memory; TAU-control = teaching as usual control.

<sup>2</sup> Due to the magnitude and duration of the study, various preliminary reports of the data were produced. These reports included different sub-samples of the study or presented analyses of the data in a variety of different ways (e.g., year by year of the study), using different data-analytic approaches, or with a variety of different software. Inevitably, there are differences between the obtained results, although all of the previous analyses have pointed to the advantage of the successful intelligence condition. This is the first presentation of the whole sample, where the analyses were carried out in the most conservative way, unit by unit, across all years of the implementation, utilizing a single analytic framework.



Table 2  
*Illustration of How Units in the Three Instructional Conditions Covered the Same Content for Students but With Different Instructional Approaches and Activities in the Language Arts Unit on Biography as a Literary Genre (1 Day)*

Successful intelligence	Critical thinking	Memory
Objectives		
Students will be able to (a) explain what a biography is, (b) identify and interpret life events, given a biographical statement, (c) compose (orally and in writing) one-sentence biographical statements.	Students will be able to (a) explain what a biography is, (b) identify and interpret life events, given a biographical statement, (c) compose (orally and in writing) one-sentence biographical statements.	Students will be able to (a) define what a biography is, (b) identify life events, given a biographical statement, (c) compose (orally and in writing) one-sentence biographical statements.
Activities		
<ul style="list-style-type: none"><li>▪ [Analytical activity]: Given a short biography, identify and categorize life events using a graphic organizer</li><li>▪ [Practical activity]: Write biographical statements about friend or family members</li></ul>	<ul style="list-style-type: none"><li>• [Analytical activity]: Given a short biography, identify and categorize life events using a graphic organizer</li><li>• [Memory activity]: Write biographical statements about the subject of a biography</li></ul>	<ul style="list-style-type: none"><li>▪ [Memory activity]: Given a short biography, recall life facts, using notes and a frame as memory aids</li><li>▪ [Memory activity]: Write biographical statements about the subject of a biography</li></ul>
Skills		
<ul style="list-style-type: none"><li>• Genre: Biography</li><li>• Description and interpretation of text</li><li>▪ Sentence writing</li></ul>	<ul style="list-style-type: none"><li>• Genre: Biography</li><li>• Description and interpretation of text</li><li>▪ Sentence writing</li></ul>	<ul style="list-style-type: none"><li>▪ Genre: Biography</li><li>• Description of text</li><li>▪ Sentence writing</li></ul>

students in the M condition had two memory activities. Table 3 provides the specific activity instructions for the classroom teacher.

**Language arts curriculum units.** Five thematic language-arts units were completed by the students in each of the three conditions (SI, CT, and M). These five units were titled (a) How and Why Nature Tales (*Wonders of Nature*); (b) Informative Nonfiction (*True Wonders*); (c) Biography (*Lively Biographies*); (d) Quest Literature (*Journeys*); and (e) Mystery (*It's a Mystery*). Thus, in total there were 15 instructionally customized newly developed units (5 content units × 3 treatment conditions). Although the content and duration of each unit were identical, within each condition, each unit was taught with different techniques based on the SI, CT, and M specifications. Students across the three conditions received the same, unit-specific pre- and posttest assessments. That is, there were five pre–posttest pairs corresponding to the five content units.

Intended as an introductory unit, *The Wonders of Nature* introduced students to two short poems about nature, which served to motivate students to “wonder” about the natural phenomena explained in *pourquoi* (“how and why”) tales. Students were taught to identify the characteristic elements of *pourquoi* tales, including the concept of cause and effect. As a culminating activity, students were expected to write their own *pourquoi* tale.

In *True Wonders*, students learned library research skills. They were expected to develop an understanding of research methods, understand the difference between fiction and nonfiction, and learn to use reading strategies to synthesize information from nonfiction sources.

In *Lively Biographies*, students were exposed to biography as a genre. They engaged in a series of activities that helped them to develop a working knowledge of the nature of the genre, the

sequencing events in chronological order, and the use of graphic organizers in the recording of events. Students then interviewed someone and produced a photo-biography.

In *Journeys*, students were engaged in the reading of quest tales and, through a series of activities, gained an understanding of the elements of the quest tale. Students were expected to articulate universal themes, identify and articulate qualities of quest heroes, and demonstrate knowledge of the above through the writing process.

Finally, in the *It's a Mystery* unit, students listened to a read-aloud mystery and at the same time independently read a mystery of their choice. Through activities based on the readings, students gained an understanding of the mystery genre, including how suspense and intrigue are built. For example, students identified the setting, characters, plot development, conflict, and resolution; learned vocabulary common to the genre; discussed human experiences and motives; and followed clues to solve the mystery. Usable data were collected for all five units (see the Note on missing data section below).

**Mathematics curriculum units.** Five mathematics units, including pre- and postintervention assessments, were completed in each of the four conditions (SI, CT, M, and TAU-control)<sup>3</sup>: (a) *Equivalent Fractions*; (b) *Measurement*; (c) *Geometry*; (d) *Data Analysis and Representation*; and (e) *Number Sense and Place Value*. Thus, in total there were 20 customized instructional units

<sup>3</sup> In our work with multiple districts and schools around the country, we established, due to the wide range of content covered in the various curricula used across the country, that the only domain in which we could implement a TAU condition was Mathematics. The diversity of curricula, pedagogies, and standards was too great in the domains of Language Arts and Science to justify a homogeneous TAU condition.

Table 3

*Detailed Descriptions of the Analytical, Practical, and Memory Activities Cited in Table 2*

Analytical activity: Listening for life facts— Biographical statement model	<p><i>After students demonstrate an understanding of biography, move on to an example of a biography. Ask the students to listen to the biography and try to identify life facts, such as date and place of birth, what the person looks like, or what the person accomplished. Write the biography on chart paper so the children can follow along while you read.</i></p> <p>Sample biography: Mrs. Murray was born in San Diego, California, and learned to swim almost before she learned to walk. Her older brother Peter taught her to swim at the marina where their dad worked as a lifeguard. As a youngster, Mrs. Murray liked to race her brother and the other children who swam at the marina. She was a tall, athletic youngster who kept her long, blond hair tied back in a ponytail. Her family was not at all surprised when she joined the high school swim team and won many medals. Today when she is not teaching her fourth grade class, Mrs. Murray still enjoys swimming and teaching her own children to swim at the local beach.</p> <p><i>Then ask the children to share “life facts” they learned about the person from hearing the brief biography. For example, they might share that Mrs. Murray is a good swimmer or that she was born in California. You might want to point out that biographies are usually written in the third person because they are about someone else’s life story. As the children share what they can remember about your life, write the “life fact” under the appropriate heading on the tag board chart. Use the category labels to prompt the children to remember life facts they heard. Tell them they can use the BIOgraphic organizer as a guide while they are reading.</i></p> <p>Note to teacher: A classroom wall chart—a BIOgraphic organizer—can be made out of tag board or flannel board for repeated use throughout the lesson. Ideally, it should be created as a pocket chart so that students can post their sentence strips to sort life facts throughout this unit. Category headings (e.g., accomplishments, appearance, family, friends, occupation) may be changed to fit reading passages. A similar matrix will be used as an advance organizer throughout the unit to assist students in reading biographies for life facts.</p>
Practical/creative activity: Writing biographical statements about friends and family	<p><i>After reading and discussing the model biography, tell the students they will finish a brief practical activity in which they will become a biographer. Tell the students that they will be doing a short activity in which they will select something memorable about a person they know well and write one biographical statement about that person. Ask the students to think of someone they know well. You may want to prompt the students to remember different aspects of the person’s life by slowly asking them a series of “remember” questions. Tell the students to close their eyes and try to remember what the person looks like, how old the person is, what the person wears, what the person likes to do, where the person works or goes to school, what friends and family members the person has, and what interesting or memorable things the person has done.</i></p> <p><i>Then ask the students to select one interesting memory and write one statement about this person. Students should write their sentences on a sentence strip/oak tag so that the sentences can be saved and referred to in future lessons, as necessary. Classroom paraprofessionals may be involved in helping the students write a complete sentence and/or checking for correct spelling and punctuation. These sentences will serve as models of short biographical statements. They will also serve as examples of “life facts” or the kinds of information a biography typically tells about a person’s life.</i></p>
Memory activity: Recall life facts	<p><i>After students demonstrate an understanding of biography, move on to an example of a biography. Ask the students to listen to the biography and take notes to memorize life facts, such as date and place of birth, what the person looks like, or what the person accomplished. Write the biography on chart paper so the children can follow along while you read.</i></p> <p>Sample biography: Mrs. Murray was born in San Diego, California, and learned to swim almost before she learned to walk. Her older brother Peter taught her to swim at the marina where their dad worked as a lifeguard. As a youngster, Mrs. Murray liked to race her brother and the other children who swam at the marina. She was a tall, athletic youngster who kept her long, blond hair tied back in a ponytail. Her family was not at all surprised when she joined the high school swim team and won many medals. Today when she is not teaching her fourth grade class, Mrs. Murray still enjoys swimming and teaching her own children to swim at the local beach.</p> <p><i>Remove the biography and ask students to recall the main life facts they just heard. Ask students to review their notes, set them aside, and then recall the main facts about the person in the biography.</i></p>

*Note.* Text in italics is for the teacher.

(5 content units  $\times$  4 treatment conditions); within each condition, each unit was taught using different techniques based on the SI, CT, M, and TAU-control specifications. However, there were only 5 pre–posttest pairs, as students across the four conditions received the same pre- and posttest assessments.

The *Equivalent Fractions* unit was intended as a follow-up to an introductory fractions unit. In it students developed an understanding of the concept of equivalence, modeled equivalent fractions with concrete manipulatives, identified and generated equivalent

fractions (denominators less than 12), and applied the concept of equivalent fractions in practical and problem-solving situations.

In the *Measurement* unit, students learned to measure quantities (including time, length, perimeter, area, weight, and volume) in everyday and problem situations. They compared, contrasted, and converted within systems of measurements (customary and metric) and estimated measurements in everyday and problem situations. In addition, students learned about the use of appropriate units and instruments for measurement.



In *Geometry*, students engaged in the identification and modeling of simple two-dimensional and three-dimensional shapes and developed an understanding of their properties (reviewing perimeter, area, and volume). Students were expected to understand and identify geometric concepts such as “congruent,” “similar,” and “symmetric.” Finally, students combined, rotated, reflected, and translated shapes.

In *Data Analysis and Representation*, students were given an opportunity to collect, organize, and display data from surveys, research, and classroom experiments. They used the concepts of range, median, and mode to describe a set of data and to interpret data in the form of charts, tables, tallies, and graphs. They learned about the use of bar graphs, pictographs, and line graphs and the advantages and disadvantages of each.

In the *Number Sense and Place Value* unit, students used number lines to identify and understand negative numbers and the ordering of numbers. They were led to an understanding of how to use the place-value structure of the Base 10 number system and how to identify factors and generate equivalent representations of numbers to use in problem solving. In addition, students explored even/odd numbers, square numbers, and prime numbers.

Usable data were collected only from three of the units: *Equivalent Fractions*, *Measurement*, and *Geometry* (see the Note on missing data section below).

**Science curriculum units.** Four science units, including pre- and postintervention assessments, were completed in each of three conditions (SI, CT, and M): (a) *The Nature of Light*; (b) *Magnetism*; (c) *Electricity*; and (d) *Ecology*. In total, there were 12 customized instructional units (4 content units  $\times$  3 treatment conditions); within each condition, each unit was taught with different techniques based on the SI, CT, and M specifications. There were only 4 pre–posttest pairs, as students across the three conditions received the same pre- and posttests.

*The Nature of Light* unit introduced the concepts of light, reflection, and refraction. By the end of this unit, students were able to show that light travels in straight lines; give examples illustrating that visible light is made of different colors; list colors of visible light; explain how a prism can separate visible light into different colors; explain how mirrors can be used to reflect light; give examples of absorption; describe and give examples of reflection; give examples and describe refraction; and describe the similarities and differences between absorption, reflection, and refraction.

In the *Magnetism* unit, students learned the properties and uses of magnets. By the end of this unit, students were able to explain the difference between magnetic and nonmagnetic objects; give examples of magnetic and nonmagnetic objects; define magnetism; predict whether two magnets will attract or repel each other; describe the effects of a magnet on a compass; explain the difference between temporary and permanent magnets; define the terms *lodestone* and *keeper* as they apply to magnetism; illustrate that the magnetic force is strongest at the poles; and identify materials that may interfere with a magnetic field.

In the *Electricity* unit, students were engaged in hands-on activities relating to electrical circuits. By the end of this unit, students were able to explain that static electricity occurs when charges are moved from one object to another; give examples of static electricity; explain how an object can become charged; define what a cell is; explain the relationship between a cell and a battery; explain what current electricity is; list the essential com-

ponents of a series circuit; explain how a series circuit works; explain how a parallel circuit works; explain the difference between a series circuit and a parallel circuit; explain what conductors are; explain what insulators are; and give examples of insulators.

In the *Ecology* unit, students were provided with a basic understanding of the interdependence of organisms and their environments through a series of activities focusing on environmental factors and their impact on animals and people, respectively, and the interdependence of plants and animals. In addition, students developed the skills necessary to conduct scientific investigations and gain an appreciation for science as a discipline. By the end of the unit, students were able to explain what a terrarium is; describe some environmental factors that are important to the growth and survival of plants and animals; give examples of how environmental factors affect the growth and survival of plants; explain how animals depend on the nonliving environment to survive; describe some environmental factors that affect animals' ability to survive and grow; give examples of the effect of the same environmental factor on different animals; describe an ecosystem; explain some of the relationships between plants, animals, and the physical environment; explain how energy passes through an ecosystem; describe the conditions that are necessary for an ecosystem to function; explain how people depend on their environment; give examples of how people can have a positive or negative effect on their environment; and understand why it is important to use natural resources wisely. Only two units, *The Nature of Light* and *Magnetism*, produced usable data (see the Note on Missing Data section below).

**Assessments.** Unit-specific assessments were developed to capture mastery in the content area of each unit but were generated in such a way that equal numbers of items tapped into the four key abilities at which the intervention conditions were aimed—that is, memory, analytical, creative, and practical abilities (Randi & Jarvin, 2006). Each pre- and posttest had 20–22 items. In order to equalize test difficulty statistically and place pre- and posttest scores on the same measurement scales, we included 3–7 items that were common to both pre- and posttest in each unit. These items were used to obtain ability scores (see below).

Initial rubrics were developed by the research team for all of the units' pre- and postintervention assessments; they were then refined in collaboration with several raters once initial student data had been collected. All the student data were then rated with the final rubrics. The items were roughly equally divided between multiple-choice (scored 0–1) and open-ended (scored 0–5) formats, with 40% to 59% identified as multiple-choice items, depending on the test. Students in all conditions received identical, unit-specific pre- and posttests. Table 4 presents the Cronbach's alpha internal consistency reliability estimates, for pre- and posttests, for both multiple-choice and open-ended questions simultaneously (Rizopoulos, 2006).

## Procedures

**Assignment to experimental groups.** Recruitment efforts were targeted at school districts rather than at individual schools, and we sought permission and buy-in from district superintendents before reaching out to principals. Depending on the size of the district and the number of schools judged by the superintendent to be candidates for

Table 4

*Internal Consistency ( $\alpha$ ) and Construct Reliability (Con.  $r_{xx}$ ) of Curriculum Unit Pretests and Posttests*

Curriculum units	Pretest			Posttest			Common items	<i>n</i>
	Items	$\alpha$	Con. $r_{xx}$	Items	$\alpha$	Con. $r_{xx}$		
Language Arts								
How and Why Nature Tales ( <i>Wonders of Nature</i> )	22	0.767	0.991	22	0.826	0.995	6	1,626
Informative Nonfiction ( <i>True Wonders</i> )	22	0.786	0.993	22	0.793	0.995	4	1,233
Biography ( <i>Lively Biographies</i> )	22	0.845	0.992	22	0.778	0.990	7	752
Quest Literature ( <i>Journeys</i> )	22	0.783	0.990	22	0.832	0.991	3	520
Mystery ( <i>It's a Mystery</i> )	22	0.813	0.992	22	0.803	0.988	7	549
Mathematics								
<i>Equivalent Fractions</i>	22	0.816	0.997	22	0.748	0.994	5	1,735
<i>Measurement</i>	22	0.698	0.992	22	0.739	0.993	3	1,550
<i>Geometry</i>	22	0.659	0.990	22	0.775	0.992	3	545
Science								
<i>The Nature of Light</i>	20	0.876	0.991	20	0.848	0.980	6	1,328
<i>Magnetism</i>	20	0.762	0.986	20	0.646	0.982	2	917

Note. Con.  $r_{xx}$  = construct reliability of factors jointly estimated with common item anchoring.

participation, one or more experimental conditions were implemented in the district. In all cases, teachers within a given school were assigned the same condition to avoid within building contamination. In other words, in small districts there might be only one participating school, so that the district is confounded with the experimental condition, whereas in a larger district, all conditions might be assigned, always to different schools. Within these constraints, the allocation to experimental condition was random. This design reflects the challenges and constraints of large-scale implementation in diverse settings, where districts and schools need to have voices in making decisions about the experimental interventions they are interested in considering. In other words, administrators decided if the district should participate, and if so, which schools should be involved, but they did not select the experimental condition(s) to be implemented. Although it is difficult to ascertain the full impact of the final allocation of schools to condition, our analyses include pretest scores as a covariate. This is in part to address concerns that even perfectly random allocation does not ensure a balance of student attributes across conditions. Yet another challenge was to get all of the participating teachers to implement all of the instructional units. Although upon recruitment districts committed to working with the whole curriculum (i.e., all units), the delivery of the full curriculum across all participating schools proved impossible due to differences between schools in terms of the content that they wanted to prioritize at the given grade level, scheduling issues due to local tests and other required activities, as well as differences among classrooms in terms of student level and speed of progression through instructional materials.

**Teacher training.** A 2-day, 12-hour in-service training program was developed and implemented by members of the research team for all the participating teachers. The workshop was tailored to the experimental condition that the participating teachers had been assigned to (i.e., SI, CT, or M).

Day 1 focused on (a) the program design, teacher requirements, and other logistics and (b) the theoretical principles of teaching and instruction for each one of the three experimental conditions. After introductions, teachers were presented with a program overview and the timeline and expectations for participation were reviewed and discussed as a group. The researchers then presented

the theoretical underpinnings and prior empirical evidence for the effectiveness of the approach (SI, CT, or M). Teachers in the SI condition thus learned about the previous studies on the effectiveness presented in the introduction to this article; teachers in the CT condition were given examples of critical thinking based instructional interventions, and teachers in the M condition were taught about the effectiveness of different mnemonic strategies for learning material. In addition to learning about earlier work, participants got to practice activities that had proven successful. Again, specific activities practiced varied between the SI, CT, and M groups. Finally, teachers practiced hands-on use of the CORE system. CORE (Collaborative Online Research Environment) is a software package that was designed specifically for this program to allow teachers to access, download, and print curriculum materials, as well as to provide a discussion board allowing them to chat both with other teachers enrolled in the same condition (SI, CT, or M) and with the curriculum developers and content specialists involved in the program.

Day 2 focused on modeling the units in each subject area and provided teachers with an opportunity for hands-on experience with the unit format. Materials distributed included a teacher guide containing instructional material, background information, resource materials reflecting print and nonprint sources, and student workbooks. Teachers received only materials relevant to the instructional approach they were to implement in their classroom. In other words, a teacher trained to implement the SI instructional approach was trained with other teachers implementing the SI approach and saw only the SI instructional materials. Teachers also were introduced to the instructional strategies particular to the condition. An overview of the pre- and postintervention assessments concluded the sessions.

**Fidelity monitoring.** Fidelity monitoring was carried out in two ways: through the CORE system and by collecting and reviewing all student workbooks to track the level of completion. As mentioned above, the CORE system is a Java-based collaborative environment designed to establish and promote long-distance collaborations with teachers. It was designed, created, and maintained by the Yale University Information Technology Department for the purposes of this program and



enabled the research team to stay in touch with implementing teachers throughout the school year. Because all electronic conversations between teachers and between teachers and research-team members were recorded and stored, the system provided data to measure fidelity of implementation. A second measure was provided by the collected student workbooks, which contained information on which part(s) of a curriculum unit and what activities had been completed by the students in a given classroom. Both teacher logs and student workbooks were analyzed for indicators of fidelity; only those teachers whose students completed all homework assignments, and whose CORE logs were indicative of both understanding of and adherence to the program, were included in the data analyses. We did not have reason to expect (and did not observe) any differences in the usage of the CORE system and in the utilization of the workbooks across instructional treatments (SI, CT, and M). In other words, there were differences across classrooms, with some but not other teachers utilizing the CORE system regularly and some teachers returning student workbooks where every activity had been completed and other teachers returning student workbooks where entire sections were blank, but these differences were observed within each instructional treatment condition. Student workbooks were used as indicators that permitted a participating classroom to be entered in the study database. If the workbook contained less than 70% of the activities completed, the data from a given teacher were not entered into the database. Altogether, ~10% of the participating classrooms in each instructional condition did not meet this criterion.

**Data processing.** All data processing was carried out at the Center for the Psychology of Abilities, Competencies, and Expertise (PACE Center) at Yale University. Details regarding the management of the data can be found in the Appendix. Close to one hundred casual employees were hired in addition to permanent research-assistant staff to assist with data entry (multiple-choice questions) and coding of open-ended questions. The open-ended questions were coded with a detailed rubric developed by the curriculum developer, and coders were trained to reach satisfactory interrater reliability levels (i.e., the correlations between the pair's open-ended item ratings had to be greater than .70) before they were allowed to start coding materials.

## Statistical Analyses

**Note on missing data.** As we worked with a large number of schools and districts, we could exercise only limited control over what and how many units were selected by teachers to be administered. Buy-in required a commitment to the whole program, but teachers needed to map their preferences for particular units onto their school calendars and other administrative demands. In turn, to include a unit into the analyses, we had to have a reasonable number of students receiving the unit across all study conditions. Unfortunately, this did not happen for two Mathematics and two Science units. Due to small or distinctly uneven distributions of the number of participants across conditions within certain units, the corresponding data were not analyzed for those four units.

**Attrition.** Extending our reporting of fidelity monitoring, a certain degree of student attrition is also expected as students come

and go throughout the school year due to illness and the like. Some students may also not be available for testing at one or the other assessment or may have joined the class part way through the training. An analysis of attrition revealed statistical differences in six of the nine units,<sup>4</sup> although effect sizes ( $\eta^2$ ) are small with no consistent pattern for any one condition. There was statistically less attrition in the SI condition for three units ( $\eta^2$ : *Equivalent Fractions* = .005; *Measurement* = .006; and *Magnetism* = .047), less attrition in the M and CT conditions for three units ( $\eta^2$ : *True Wonders* = .042; *The Wonders of Nature* = .015, and *mysteries* .028), and no statistical differences in attrition for the remaining three units. Of importance, these differences were not related to the intervention differences to be reported shortly. Only students who were available for assessment at both time points were included in the analyses.

**Overview of analyses.** The analyses we report here were conducted in two stages. First, we derived performance measures for each unit. Second, we ran unit-specific analyses that included a set of covariate and interaction terms.<sup>5</sup> The rationale and general approach for these are described next.

**Derivation of performance measures.** To combine multiple-choice (binary) and open-ended (ordinal) items into a single ability score, we used Samejima's graded response model (Samejima, 1997), as implemented in Mplus (Muthén & Muthén, 2005), for both pre- and posttest data (such scores have a range of approximately -3 to 3). For the overlapping items that were presented both at pre- and posttest, their loading and threshold (i.e., their discrimination and difficulty parameters) were constrained. This allowed for the statistical equating of pre- and posttest item difficulty. As recommended in the literature (Geiser, Eid, Nussbeck, Courvoisier, & Cole, 2010), scores were calculated for only those individuals with both pre- and posttest data. We do not elaborate on these analyses here; however, details can be obtained from the authors. Traditional internal consistency measures for the pre- and posttest assessments of each unit are provided in Table 4, along with construct reliability estimates (Gefen, Straub, & Boudreau, 2000).

**Unit-specific analyses with covariates.** As we have described, the students who participated in the current study were sampled from a large and diverse population. One of the touted benefits of a cognitive approach to educational interventions is the real possibility of capturing a much broader and diverse range of approaches to learning. This has certainly been our general experience in the smaller scaled applications of the theory of successful intelligence. One difficulty we faced in the current study is that student-level diversity (e.g., gender and ethnicity) was not collected for reasons described previously. We attempt to capture this diversity and the differential extent that it may impact performance across condition by using a number of school- and classroom-level covariates. The diversity we are capturing is thus in terms of the educational environment, not the child's specific circumstances.

<sup>4</sup> Attrition here is defined as data not available at either pretest or posttest.

<sup>5</sup> These analyses are the culmination of a comprehensive series of analytics conducted in a number of passes across this large database. We acknowledge the reviewers' significant input in shaping the final set we report here.

Following the derivation of measures for each educational unit (5 Language Arts, 3 Mathematics, and 2 Science units), a series of mixed-effects (multilevel) regressions was fit to estimate the effect of intervention condition on the posttest performance. The pretest was always included as a covariate in the regressions. To evaluate the robustness of the obtained results, we repeated the analyses using, *inter alia*, alternative centering (group mean), a different random clustering variable (school rather than teacher), and the propensity scores approach to match the experimental groups as closely as possible (Dehejia & Wahba, 2002; Ho, Imai, King, & Stuart, 2007). Although there was some variability in the findings (i.e., the magnitude of effects), the pattern of results was generally consistent.<sup>6</sup> The approach we used for the analyses reported here is as follows: There were two levels in the multilevel analysis: students at Level 1 clustered within classroom teachers at Level 2. That is, random effects (covariates and intervention conditions) were estimated at the teacher level (Level 2) to account for classroom level clustering. Students' posttest and pretest performances were modeled at Level 1. Where statistically possible, all models included critical classroom- and school-level demographic variables and their interaction with experimental condition. *Title I status*,<sup>7</sup> *gender* (defined as the proportion of the school population that was male; i.e., % male) and % White (proportion of the school population that was White) were school-level variables, and *giftedness* (whether the class was identified as a regular or gifted-education classroom) was a classroom-level variable. The % male and % White variables were grand-mean centered for entry alone and as part of interaction terms. It is conceivable to introduce school variability as a third level in the model by clustering classrooms within schools. However, the distribution of the number of classes across schools and intervention conditions was quite broad—on average there were only 1.63 classrooms per school (standard deviation = .45). This suggested to us (and was supported by our preliminary analyses) that the school-level variables would provide little additional statistical information (in relation to their association with student performance) if they were modeled at the school level, rather than at the classroom/teacher level. Furthermore, given the limited variability in number of classrooms per school, a three-level model would be unstable. As such, the decision was made to stay with the simpler two-level model. The regression models were fit in R with the nonlinear mixed effects models (NLME) package (Pinheiro, Bates, DebRoy, Sarkar, & the R Core team, 2009). Note that the NLME package accommodates both linear and nonlinear models; however, in the present study only linear models were run. We treated intervention conditions as multiple, dummy coded variables (with SI as the reference group) in the analyses for each unit. In one or more conditions of some units, covariates were constants or zero. They were excluded from analyses when this occurred.

## Results

### Sample Data

Table 5 presents descriptive statistics of the unadjusted pretest and posttest performance scores, and the characteristics of the sample by study condition. Of note is the large variability in sample characteristics among the different conditions and different units. This reflects the realities of conducting research

during real-time classroom teaching using intact classrooms. To control statistically for this variability, we fit regressions separately for each unit and included pretest as a Level 1 covariate and demographic variables (Title I, % male, % White, and giftedness) as Level 2 covariates. Interaction terms were also entered when possible to capture (in part) variability in the differential functioning of covariates between conditions. All models were run with the same set of covariates first, and for those models that would not statistically converge with all covariates, the models were modified. Covariates not able to be included for a particular model are represented with a dash in Table 6. Regressions were fit with varying intercepts and were grand-mean centered (*Title I* and *giftedness* indicators were not centered because these are binary variables). The analyses, which included the intervention condition coded into multiple dummy-variables with SI as the reference group (i.e., CT vs. SI, and M vs. SI, and, in addition for Mathematics units, TAU vs. SI), revealed the following results. First, all unit analyses included the student-level pretest score as a covariate, and in all cases, as would be expected, it was a statistically significant predictor of posttest performance. We report unstandardized regression coefficients in Table 5 and the graphical representation of this data in Figure 1 (along with 95% confidence intervals). Below is a summary of the results for each academic domain.

### Units

**Language arts units.** There were five language arts units that had analyzable data. Three of the five had a statistically significant effect for intervention condition. Controlling for student pretest score and school-level covariates (gender, % White, and Title I, and their interaction with condition) there was a statistically significant advantage to the SI condition over the CT condition in *Wonders of Nature* ( $b = -0.86, p = .05$ ) and *Journeys* ( $b = -0.29, p = .02$ ). CT was superior to SI in *Mysteries* ( $b = 0.81, p = .01$ ). There were no statistically significant intervention effects for any of the other Language Arts units.

**Mathematics units.** Three mathematics units had analyzable data. Two of the three had statistically significant effect for intervention condition. Controlling for pretest performance and Level 2 covariates, statistically significant intervention effects were observed for *Equivalent Fractions* in favor of SI over TAU ( $b = -0.27, p = .01$ ) and for *Measurement* in favor of Memory over the SI intervention ( $b = 0.28, p < .04$ ). There were no statistically significant intervention effects for *Geometry*.

**Science units.** There were two science units that had analyzable data, and both had statistically significant effect for intervention condition. For *The Nature of Light* unit, there was a significant intervention effect in favor of SI over Memory ( $b = -0.78, p < .01$ ). For *Magnetism*, there was a significant

<sup>6</sup> All of these results, as well as the details of the results presented in this article, are available from the authors upon request.

<sup>7</sup> We used Title I data (<http://nces.ed.gov/>) for each school as a proxy for socioeconomic status.



Table 5  
Descriptive Characteristics of the Study Groups

Curriculum units	Test results <sup>a</sup>				N	Demographic characteristics <sup>b</sup>			
	Pretest		Posttest			% girls	% White	Title I	Giftedness
	M	SD	M	SD					
Language arts									
How and Why Nature Tales ( <i>Wonders of Nature</i> )									
SI	0.01	0.87	0.54	1.35	703	51.1	63.2	48.1	26.9
CT	−0.02	0.87	0.43	1.13	542	47.5	88.2	30.8	39.1
M	0.18	0.97	−0.02	1.30	436	49.1	63.6	88.1	24.5
Informative Nonfiction ( <i>True Wonders</i> )									
SI	0.03	1.03	0.11	0.84	519	50.2	73.2	31.4	34.9
CT	−0.08	0.69	0.02	0.63	377	47.8	88.2	29.2	34.0
M	−0.24	0.95	−0.09	0.81	337	49.2	64.6	82.8	28.2
Biography ( <i>Lively Biographies</i> )									
SI	−0.09	0.98	0.00	0.41	340	53.6	72.4	69.7	0.0
CT	−0.09	0.87	−0.04	0.43	220	48.2	79.7	56.4	0.0
M	−0.20	0.91	−0.02	0.39	192	48.8	59.1	100.0	0.0
Quest Literature ( <i>Journeys</i> )									
SI	0.03	0.91	0.20	0.82	322	55.0	68.8	56.8	0.0
CT	−0.25	1.04	−0.25	1.05	144	49.2	75.3	45.1	0.0
M	−0.11	0.72	0.12	0.74	89	52.5	83.8	100.0	0.0
Mystery ( <i>It's a Mystery</i> )							100.0		
SI	−0.16	0.75	0.08	0.41	232	52.1	62.9	90.5	0.0
CT	−0.59	1.15	−0.05	0.42	157	48.8	88.2	32.5	0.0
M	−0.31	0.68	−0.02	1.30	160	50.0	76.9	100.0	0.0
Mathematics									
Equivalent Fractions									
SI	−0.19	0.89	−0.06	0.46	663	50.5	74.8	24.1	57.5
CT	−0.31	0.94	−0.06	0.47	585	48.8	67.9	21.4	65.5
M	−0.40	0.81	−0.03	0.43	451	50.3	74.5	36.8	47.5
TAU	−1.09	0.57	−0.70	0.34	36	47.9	70.2	100.0	0.0
Measurement									
SI	0.16	0.81	0.09	0.92	548	50.4	78.0	19.0	59.7
CT	−0.03	0.95	0.02	1.04	485	48.3	77.4	27.2	67.0
M	−0.12	0.86	0.01	0.92	485	49.8	69.4	45.4	47.2
TAU	−0.85	0.93	−0.89	1.09	32	47.9	70.2	100.0	0.0
Geometry									
SI	−0.24	0.89	−0.20	0.69	284	50.1	54.1	68.7	0.0
CT	0.65	0.68	0.65	0.55	128	50.1	80.2	4.7	100.0
M	−0.10	0.69	−0.07	0.68	103	47.7	67.2	100.0	0.0
TAU	−1.03	0.60	−0.56	0.75	30	47.9	70.2	100.0	0.0
Science									
The Nature of Light									
SI	0.09	0.94	0.01	0.31	617	49.9	69.7	20.5	76.3
CT	0.17	0.86	0.08	0.34	444	49.1	72.7	5.6	81.3
M	−0.36	0.84	−0.16	0.27	267	47.3	62.7	30	63.7
Magnetism									
SI	0.05	0.83	−0.08	0.72	345	52.5	65.4	0.0	84.6
CT	−0.24	0.86	0.18	0.60	453	47.7	69.4	0.0	100.0
M	−0.38	0.98	0.03	0.57	119	47	79.7	0.0	100.0

Note. SI = successful intelligence; CT = critical thinking; M = memory; TAU = teaching as usual control.

<sup>a</sup> The pretest and posttest scale is a function of Samejima's graded response model (Samejima, 1997); 0 is defined as the average ability level for individuals as measured by the test. <sup>b</sup> School-level data (average of the percentage of students in the school for a given characteristic).

advantage for the critical thinking condition over SI ( $b = 0.32$ ,  $p = .04$ ).

### Summary of Analyses

In sum, the analyses, which included the intervention condition coded into multiple dummy-variables with SI as the reference group, revealed 7 effects (out of 23) of mention. There were four

cases where SI was advantageous (*Wonders of Nature*, *Journeys*, *Equivalent Fractions*, and *Light*), one case where Memory was advantageous (*Measurement*), and two cases in favor of Critical Thinking (*Mysteries* and *Magnetism*). This is not substantially different from what we might expect by chance. The SI intervention did not lead to an overall advantage as expected, but equally it did not lead to a disadvantage.

Table 6  
Regression Coefficients (and *p* Values in Parentheses) for Multilevel Analyses of Curriculum Units

Effects	Wonders of Nature	True Wonders	Lively Biographies	Journeys	It's a Mystery	Equivalent Fractions	Measurement	Geometry	The Nature of Light	Magnetism
Conditions vs. SI										
CT	<b>-0.86 (0.05)</b>	0.09 (0.68)	-0.06 (0.80)	<b>-0.29 (0.02)</b>	<b>0.81 (0.01)</b>	-0.02 (0.84)	0.08 (0.59)	0.23 (0.20)	-0.11 (0.55)	<b>0.32 (0.04)</b>
M	-0.68 (0.21)	0.32 (0.39)	0.11 (0.14)	0.48 (0.26)	-0.11 (0.34)	0.06 (0.66)	<b>0.28 (0.04)</b>	-0.02 (0.93)	<b>-0.78 (0.00)</b>	0.15 (0.46)
TAU	—	—	—	—	—	<b>-0.27 (0.01)</b>	0.12 (0.30)	0.21 (0.15)	—	—
Covariates										
Pretest	<b>1.01 (0.00)</b>	<b>0.64 (0.00)</b>	<b>0.35 (0.00)</b>	<b>0.82 (0.00)</b>	<b>0.33 (0.00)</b>	<b>0.40 (0.00)</b>	<b>0.94 (0.00)</b>	<b>0.70 (0.00)</b>	<b>0.21 (0.00)</b>	<b>0.42 (0.00)</b>
% White	<b>0.30 (0.00)</b>	0.12 (0.17)	-0.06 (0.28)	-0.04 (0.63)	0.02 (0.72)	<b>0.14 (0.00)</b>	0.06 (0.29)	-0.05 (0.62)	-0.04 (0.55)	<b>0.17 (0.01)</b>
% male	0.02 (0.77)	0.03 (0.65)	-0.05 (0.10)	0.03 (0.61)	0.03 (0.47)	-0.06 (0.21)	0.01 (0.83)	0.04 (0.70)	<b>-0.07 (0.03)</b>	-0.01 (0.94)
Title I	-0.10 (0.71)	0.18 (0.32)	0.01 (0.85)	-0.03 (0.80)	<b>0.52 (0.00)</b>	0.06 (0.47)	0.00 (0.98)	0.12 (0.48)	-0.16 (0.16)	—
Giftedness	0.18 (0.54)	0.26 (0.18)	—	—	—	-0.04 (0.68)	<b>0.35 (0.01)</b>	—	-0.17 (0.32)	—
Interactions										
CT × % White	0.32 (0.46)	-0.23 (0.40)	0.12 (0.55)	0.10 (0.27)	-0.25 (0.27)	-0.09 (0.14)	-0.12 (0.06)	0.19 (0.55)	0.01 (0.93)	—
M × % White	<b>-0.50 (0.00)</b>	-0.14 (0.19)	0.09 (0.16)	-0.99 (0.34)	-0.11 (0.34)	-0.04 (0.77)	-0.07 (0.42)	0.12 (0.53)	<b>0.27 (0.01)</b>	—
CT × % Male	0.25 (0.25)	-0.12 (0.55)	0.08 (0.87)	—	<b>-0.99 (0.05)</b>	0.07 (0.25)	0.08 (0.56)	—	0.11 (0.06)	—
M × % Male	-0.09 (0.69)	-0.04 (0.77)	-0.10 (0.48)	—	-0.05 (0.79)	0.03 (0.53)	-0.08 (0.29)	—	<b>1.01 (0.00)</b>	—
CT × Title I	0.07 (0.83)	-0.11 (0.67)	—	—	—	0.13 (0.34)	-0.18 (0.37)	—	<b>0.40 (0.04)</b>	—
M × Title I	0.52 (0.37)	-0.46 (0.24)	—	—	—	-0.12 (0.33)	-0.06 (0.67)	—	<b>0.81 (0.01)</b>	—
CT × Giftedness	0.11 (0.80)	0.14 (0.63)	—	—	—	0.13 (0.31)	-0.07 (0.69)	—	0.16 (0.49)	—
M × Giftedness	0.19 (0.63)	-0.07 (0.80)	—	—	—	0.24 (0.31)	-0.18 (0.35)	—	0.02 (0.92)	—
Intercept	<b>0.66 (0.00)</b>	-0.07 (0.53)	<b>0.66 (0.00)</b>	<b>0.19 (0.01)</b>	<b>-0.31 (0.03)</b>	-0.02 (0.78)	<b>-0.22 (0.01)</b>	-0.12 (0.38)	0.14 (0.32)	-0.08 (0.43)

Note. Dependent variable is the scaled posttest performance score. Details of covariates and analyses are provided in the text. Bold values represent results that are statistically significant at  $p < .05$ . Dashes represent covariates that were not able to be included for a model. SI = successful intelligence; CT = critical thinking; M = memory; TAU = teaching as usual.



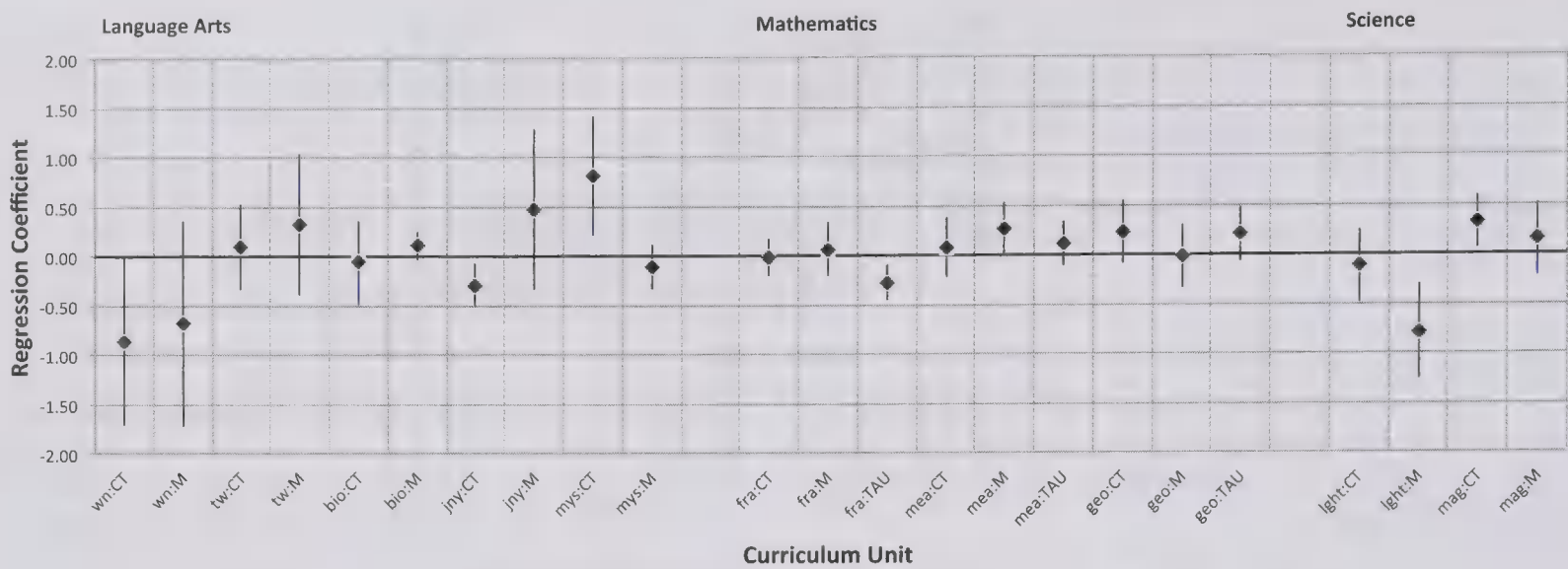


Figure 1. Regression coefficients and 95% confidence intervals of students in the SI condition relative to experimental conditions for each curriculum unit. The SI group is set at 0.00. Correspondingly, all units with coefficients below the 0 line indicate an advantage of the SI condition (and conversely for coefficients above the 0 line). Conditions: SI = successful intelligence; CT = critical thinking; M = memory; TAU = teaching as usual. Language Arts units: wn = *Wonders of Nature*; tw = *True Wonders*; bio = *Lively Biographies*; jny = *Journeys*; mys = *It's a Mystery*. Mathematics units: fra = *Equivalent Fractions*; mea = *Measurement*; geo = *Geometry*. Science units: lght = *The Nature of Light*; mag = *Magnetism*.

The pattern of influence of the covariates, both alone and as interactions, is varied across interventions (see Table 6, covariates). This pattern attests to the diversity of variables that influence, in complex ways, attempts to scale experimental investigations of intervention effects into everyday contexts. Controlling for these demographic characteristics of the schools and classrooms using the data we have access to, the SI intervention was advantageous in each domain (Language Arts, Mathematics, and Science) but weakly and inconsistently so.

Discussion

Based on the data collected in previous studies and discussed in the introduction, teaching for successful intelligence has been shown to help improve instruction and assessment in a variety of disciplines at diverse grade levels (Grigorenko et al., 2002; Sternberg et al., 1998, 2011). Most important, SI research has helped to provide a way of showing that if students are taught in a way that fits their ability profiles, they will achieve at higher levels and be better able to leverage their diverse skills (Sternberg et al., 1999).

The range of results found in the current study across all units and conditions are discordant with our previous findings. That is, regardless of (a) the rigorous research design, (b) the substantial resources invested by our team of highly skilled researchers drawn from around the world, as well as the numerous classroom teachers who invested time and energy to be involved, (c) the infrastructure available from one of the very best universities in the world in which the project was hosted, and of course (d) the recognition and support of the National Science Foundation (NSF) granting committee who invested in the SI theory to fund this large-scale research project, the results are sobering, especially in light of our previous successes. Because of the investments of the many stakeholders involved with the

project, it is incumbent on us to reflect on the implications of these findings in relation to the future of SI theorizing and for educational research that aims to scale up interventions that have previously demonstrated advantages in small, controlled studies. In this regard we first consider the future utility of the “economy of scale” argument, on which large-scale intervention studies are often grounded, and second reflect on the specific implications of scaling the SI intervention relative to the strong control interventions in regard to implementation fidelity.

Economies of Scale: Is It a Viable Approach?

One potential explanation for the observed results is that the attempt to apply economic theories and models to education may be fundamentally misguided. Many policymakers endorse a factory metaphor for thinking about education, in which students are the “products” to be filled with knowledge and teachers are a means of production (see Madaus, Haney, & Kreitzer, 1992, for a description). The microeconomics concept of economies of scale, upon which the notion of scaling up educational interventions rests, has been demonstrated to be highly effective in the manufacturing world (e.g., Henry Ford’s assembly line). However, Seddon (2010) has argued convincingly that economies of scale are not applicable in the context of human service professions, and educational delivery is arguably much more closely aligned with human services than with a factory metaphor. Further, as Elias et al. (2003) noted, one of the reasons that scaling up educational interventions is challenging is because educational interventions primarily rely on human operators rather than technologies. Teachers are not automatons that execute a standardized curriculum in a standardized way. Rather, Elias et al. suggest that a more useful metaphor for thinking about scaling up interventions is a sailing analogy in which various elements of the environment can take a

toll on a successful voyage and thus call to the forefront the skill of the sailors in navigating the environment. In addition, given the long history of local control of education in the United States, each state, district, and even school may have a unique cultural, organizational, and educational context (Stemler & Bebell, 2012). Although there is currently a movement toward the development of Common Core Standards in education in an effort to reduce some of the variability in curricular issues, this will not address all of the systemic variability that can impact efforts to scale up educational interventions.

Given the rigor strived for but not necessarily fully attained in the current investigation, our data suggest that it may be time to abandon the illusion that economies of scale should be pursued in the context of educational interventions. Instead, alternate models such as those being embraced by various teacher education programs throughout the country currently appear to us to be more promising. These models take a very different approach in which the implementation is tightly monitored and supported and in which new organizations wishing to join must be evaluated for the relevance of their contextual characteristics.

### Implementation Fidelity at Scale: SI Dynamics

Traditional higher level teaching interventions, like training for memory skills, are formidable interventions against which to pit new teaching approaches for a number of reasons. First, traditional, memory-based strategies are the ones teachers may be expected to revert to in uncertain situations (e.g., when attempting to implement a new teaching philosophy for the first time). It takes time for teachers to acclimate themselves to a new philosophy, and two days of teacher training, although the most we could request, simply may not be sufficient. Second, given that the SI condition includes traditional memory and critical thinking aspects, as well as creative and practical ones, it may be possible for teachers to focus on more traditional aspects and still feel they are appropriately adhering to the SI condition. Third, it is important to remember that the unit content was identical across all conditions. The differences between intervention conditions were in the framing of the teacher training, which included differential instruction in the underlying philosophy of SI, M, or CT, as appropriate. Furthermore, the curriculum content across all units and conditions was strong and well structured enough to provide engaging activities aimed at facilitating knowledge acquisition in the specific domain regardless of the intervention framing. Fourth, just as it is expected to take time for teachers to acclimate themselves to the SI philosophy, students also need time to adjust to differences in instruction (Jeltova et al., 2011). Finally, many of the content areas chosen for the units inherently required analytic skills and the memorization of facts. This is certainly true for the Mathematics units and to a lesser extent the Science units. However, it is also true of the Language Arts units.

It also is possible either that the SI model does not work effectively for all the conditions we studied or that our realization of it was less than fully effective. It would take further research to elicit a more definitive answer to such questions.

### Limitations

A study such as this one obviously has its limitations. We consider population issues, cost-benefit issues, and teacher and student issues that impact on fidelity of implementations.

**Population issues.** All students were fourth graders, and only three academic subjects were used. The sheer scale of the study practically ensured that some implementation sites would have higher fidelity than others. In addition, given that the study unfolded in nine states across the country, it was impossible to utilize a single standardized achievement test across all study groups and all domains. A measure of overall achievement (i.e., an end-of-year standardized achievement test) would have provided an alternative test of effectiveness.

**Cost-benefit analyses.** As innovation and change are costly, a fair question to consider is the cost-benefit analyses that compare the obtained gain in achievement to the costs of introducing a change in instruction. This question has not been the focus of investigation in studies introducing cognitive theories of learning-based approaches to classrooms, and the theory of successful intelligence is not to be excluded. Nevertheless, we are not prepared to conclude just yet that cognitive-based interventions, including those grounded in the theory of successful intelligence, generally do not lead to sufficient enhanced student achievement to be worth the effort. This is in part because the specific advantages of cognitively based interventions may interact with content, school-level variables and the scale of the implementation in complex and dynamic ways.

Insights from the present efforts to upscale an instructional intervention within the context of an experimental study are consistent with those stated in the literature. First, teacher buy-in plays a critical role in the success of any curricular intervention. Throughout the year, teachers inevitably faced many external demands that compromised their ability to complete all of the intended units. Second, when working with intact classrooms, there are potential confounds that can creep into study design. In the case of the current study, there are examples in which the instructional condition is confounded with a particular type of classroom (e.g., gifted classrooms that received memory-based instruction). Such anomalies cannot be co-varied out.

**Teacher and student issues.** We found perhaps the most challenging aspect of the study to be teachers' differential comfort levels with various instructional methods. Even though teachers were trained in the teaching method they would use, when under stress, we might expect some teachers to revert to what is easiest and most familiar. Under the pressures of day-to-day teaching over the long term, which poses different demands than either teaching for a laboratory experiment or teaching for a short-term study, even teachers who are well trained in a new method may find themselves reverting to older, more familiar methods that they can use without the constant vigilance and concentration required of new interventions. They revert because they are under so many other pressures: classroom management issues, parental pressures, and administrative mandates that they need to confront at the same time. Fidelity to treatment method thus becomes an issue, and such violations of fidelity are particularly difficult to control in the context of a large-scale study such as ours.

A further issue is students' own comfort with different methods of instruction. Students, like teachers, are simply much more familiar with memory-based instruction than with other methods used in the teaching/learning process. Because the students' mental



resources often are split between listening to the teacher, thinking about and planning for events going on in their extracurricular lives, and engaging in the social context of the classroom, they as well may find it easier to relate to traditional teaching than to novel methods of instruction.

## Summary

In sum, the results of this large-scale, multistate study suggest that there are difficulties associated with scaling up educational interventions that have been demonstrated to be effective in smaller contexts. Implementation of the curricular materials was designed and implemented with a minimal level of support from the research team, and the student achievement results revealed that the impact of the curriculum on student performance, when compared with strong pedagogical approaches involving teaching for memory and/or critical thinking, as well as with “teaching-as-usual” approaches, was heterogeneous. The results suggest that SI instruction does lead to student achievement outcomes that are, at a minimum, generally equivalent to those associated with other strong instructional interventions. Overall, the effects were weak, and the pattern of influence of the school and classroom covariates on posttest performance differed across interventions and units. Across the domains of literature, mathematics, and science, enhanced student performance was observed in only 7 out of 23 comparisons. SI was advantageous in four cases. There was one case where M was advantageous and two cases in favor of CT.

The traditional approach would be to conduct more rigorous, lab-like investigations into SI effectiveness; consequently, smaller replications of this study in different contexts might be called for. Or, it might be suggested that we investigate our critical thinking and memory interventions more rigorously. However, it is important to recognize that such rigor, by definition, introduces into the investigation constraints that are not feasible in real, intact classrooms—constraints we specifically set out to free in the current study.

It is important to place the data, results, and related discourse presented here in the larger context of the relevant literatures and question whether we as a research group, and the discipline in general, are going about such investigations the wrong way. Should implementation of interventions be tightly monitored and supported and participation eligibility be evaluated for relevance of contextual characteristics? These questions need deep reflection. The following observations seem to be important.

First, even if a particular instructional approach has generated robust evidence pertaining to its efficacy and replication, this does not mean that the scaling it up will be as effective as its more controlled, smaller scale evaluations. We argue that such a diffusion of the promise of an intervention is linked, primarily, to contextual factors, both systematic and random, influencing the context in which the intervention is scaled. This observation is relevant not only to the work presented here but to many other educational interventions. Second, it appears that scaled-up interventions may be characterized by a decrease of effect sizes observed in more controlled evaluations of the efficacy and robustness of an experimental pedagogy. Third, systematic efforts are needed (a) to characterize and parameterize contextual factors that threaten the consistency of an intervention when scaling up and (b) to quantify the expected decrease on previously reported intervention effect sizes. These

issues should be factored into the cost–benefit analyses of implementing change in education and should inform policy decision making. In such analyses and decisions, the empirical challenges to an innovation should be considered along with the humanistic and societal values and the ever-changing demands of the labor market. Factors such as these often do not wait for the relevant rigorous studies to be completed in a time comparable to the dynamics of real life.

## References

- Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2009). *Memory*. New York, NY: Psychology Press.
- Bruning, R. H., Schraw, G. J., & Norby, M. M. (2010). *Cognitive psychology and instruction* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.
- Clements, D. (2005). Scaling up TRIAD: Teaching early mathematics for understanding with trajectories and technologies [Grant proposal]. Retrieved from the Institute for Educational Sciences website: <http://ies.ed.gov/funding/grantsearch>
- Constas, M., & Sternberg, R. J. (Eds.). (2006). *Translating educational theory and research into practice*. Mahwah, NJ: Erlbaum.
- Corno, L., Cronbach, L. J., Kupermintz, H., & Lohman, D. F. (2001). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. New York, NY: Routledge.
- Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York, NY: Irvington.
- Dehejia, R., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84, 151–161. doi:10.1162/003465302317331982
- Elias, M. J., Zins, J. E., Graczyk, P. A., & Weissberg, R. P. (2003). Implementation, sustainability, and scaling up of social-emotional and academic innovations in public schools. *School Psychology Review*, 32, 303–319.
- Elmore, R. (1996). Getting to scale with good educational practice. *Harvard Educational Review*, 66, 1–26.
- Folland, S., Goodman, A., & Stano, M. (2013). *The economics of health and health care* (7th ed.). New York, NY: Pearson.
- Francis, D. (2011). Scale-up evaluation of reading intervention for first grade English learners [Grant proposal]. Retrieved from the Institute for Educational Sciences website: <http://ies.ed.gov/funding/grantsearch>
- Fuchs, D. (2004). Scaling up peer assisted learning strategies to strengthen reading achievement [Grant proposal]. Retrieved from the Institute for Educational Sciences website: <http://ies.ed.gov/funding/grantsearch>
- Gardner, H. (1993). *Multiple intelligences: The theory in practice*. New York, NY: Basic Books.
- Gardner, H. (2006). *Multiple intelligences: New horizons in theory and practice*. New York, NY: Basic Books.
- Gefen, D., Straub, D., & Boudreau, M. (2000). Structural equation modeling and regression: Guidelines for research practice. *Communications of the Association for Information Services*, 1(7), 1–78.
- Geiser, C., Eid, M., Nussbeck, F. W., Courvoisier, D. S., & Cole, D. A. (2010). Analyzing true change in longitudinal multitrait-multimethod studies: Application of a multimethod change model to depression and anxiety in children. *Developmental Psychology*, 46, 29–45. doi:10.1037/a0017888
- Glennan, T. K., Bodily, S. J., Galegher, J. R., & Kerr, K. A. (2004). *Expanding the reach of educational reforms: Perspectives from leaders in the scale-up of educational interventions*. Santa Monica, CA: RAND Corporation.
- Grigorenko, E. L., Jarvin, L., Diffley, R., Goodyear, J., Shanahan, E. J., & Sternberg, R. J. (2009). Are SSATs and GPA enough? A theory-based approach to predicting academic success in secondary school. *Journal of Educational Psychology*, 101, 964–981. doi:10.1037/a0015906

- Grigorenko, E. L., Jarvin, L., & Sternberg, R. J. (2002). School-based tests of the triarchic theory of intelligence: Three settings, three samples, three syllabi. *Contemporary Educational Psychology*, 27, 167–208. doi:10.1006/ceps.2001.1087
- Grigorenko, E. L., & Sternberg, R. J. (2001). Analytical, creative, and practical intelligence as predictors of self-reported adaptive functioning: A case study in Russia. *Intelligence*, 29, 57–73.
- Halpern, D. F. (1996). *Thought and knowledge: An introduction to critical thinking*. Mahwah, NJ: Erlbaum.
- Ho, D., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15, 199–236. doi:10.1093/pan/ mpl013
- Hurtig, R. (2004). Breakthrough to literacy in the Chicago public schools: A large-scale evaluation of the effectiveness of a reading comprehension interventions [Grant proposal]. Retrieved from the Institute for Educational Sciences website: <http://ies.ed.gov/funding/grantsearch>
- Jeltova, I., Birney, D., Fredine, N., Jarvin, L., Sternberg, R. J., & Grigorenko, E. L. (2011). Making instruction and assessment responsive to diverse students' progress: Group-administered dynamic assessment in teaching mathematics. *Journal of Learning Disabilities*, 44, 381–395. doi:10.1177/0022219411407868
- Kornilov, S. A., Tan, M., Elliott, J. G., Sternberg, R. J., & Grigorenko, E. L. (2012). Gifted identification with Aurora: Widening the spotlight. *Journal of Psychoeducational Assessment*, 30, 117–133. doi:10.1177/0734282911428199
- Madaus, G. F., Haney, W., & Kreitzer, A. (1992). *Testing and evaluation: Learning from the projects we fund. Policy issues in the conduct of corporate support for education*. Washington, DC: Council for Aid to Education.
- Mayer, R. E. (2011). Intelligence and achievement. In R. J. Sternberg & S. B. Kaufman (Eds.), *Cambridge handbook of intelligence* (pp. 738–747). New York, NY: Cambridge University Press.
- McKenna, M. C., & Walpole, S. (2010). Planning and evaluating change at scale: Lessons from Reading First. *Educational Researcher*, 39, 478–483. doi:10.3102/0013189X10378399
- Muthén, L. K., & Muthén, B. O. (2005). *Mplus user's guide*. Los Angeles, CA: Author.
- Nunnery, J. A. (1998). Reform ideology and the locus of development problem in educational restructuring: Enduring lessons from studies of educational innovation. *Education and Urban Society*, 30, 277–295. doi:10.1177/0013124598030003002
- Ormrod, J. E. (2010). *Educational psychology: Developing learners* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.
- Pane, J. (2007). Effectiveness of cognitive tutor Algebra One implemented at scale [Grant proposal]. Retrieved from the Institute for Educational Sciences website: <http://ies.ed.gov/funding/grantsearch>
- Pashler, H., McDaniell, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest*, 9, 106–119.
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & the R Core team. (2009). nlme: Linear and nonlinear mixed effects models. R package version 3.1-92 [Computer software].
- Randi, J., & Jarvin, L. (2006). An “A” for creativity: Assessing creativity in the classroom. *The Thinking Classroom*, 7(4), 26–32.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory. *Journal of Statistical Software*, 17, 1–25.
- Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York, NY: Springer.
- Seddon, J. (2010). *Why do we believe in economies of scale?* Retrieved from <http://www.vanguard-consult.dk/wp-content/uploads/2011/10/whydowebeleieveineconomiesofscale.pdf>
- Slavin, R. E. (2008). *Educational psychology: Theory and practice*. Needham Heights, MA: Allyn-Bacon.
- Spear, L. C., & Sternberg, R. J. (1987). Teaching styles: Staff development for teaching thinking. *Journal of Staff Development*, 8, 35–39.
- Starkey, P. (2004). Scaling up the implementation of a pre-kindergarten mathematics curriculum in public preschool programs [Grant proposal]. Retrieved from the Institute for Educational Sciences website: <http://ies.ed.gov/funding/grantsearch>
- Stemler, S. E., & Bebell, D. (2012). *The school mission statement: Values, goals, and identities in American education*. Larchmont, NY: Eye on Education.
- Stemler, S. E., Grigorenko, E. L., Jarvin, L., & Sternberg, R. J. (2006). Using the theory of successful intelligence as a basis for augmenting AP exams in psychology and statistics. *Contemporary Educational Psychology*, 31, 344–376. doi:10.1016/j.cedpsych.2005.11.001
- Stemler, S. E., Sternberg, R. J., Grigorenko, E. L., Jarvin, L., & Sharpes, D. K. (2009). Using the theory of successful intelligence as a framework for developing assessments in AP Physics. *Contemporary Educational Psychology*, 34, 195–209. doi:10.1016/j.cedpsych.2009.04.001
- Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York, NY: Cambridge University Press.
- Sternberg, R. J. (1995). *In search of the human mind*. Orlando, FL: Harcourt Brace College.
- Sternberg, R. J. (1997). *Successful intelligence*. New York, NY: Plume.
- Sternberg, R. J. (2003a). *Sternberg Triarchic Abilities Test*. Unpublished manuscript, Yale University.
- Sternberg, R. J. (2003b). *Wisdom, intelligence, and creativity synthesized*. New York, NY: Cambridge University Press.
- Sternberg, R. J. (2005). The theory of successful intelligence. *Revista Interamericana de Psicología*, 39, 189–202.
- Sternberg, R. J. (2009). WICS: A new model for liberal education. *Liberal Education*, 95(4), 20–25.
- Sternberg, R. J. (2010). *College admissions for the twenty-first century*. Cambridge, MA: Harvard University Press.
- Sternberg, R. J., Birney, D., Jarvin, L., Kirlik, A., Stemler, S., & Grigorenko, E. L. (2006). From molehill to mountain: The process of scaling up educational interventions (First-hand experience upscaling the theory of successful intelligence). In M. Constanas & R. J. Sternberg (Eds.), *Translating educational theory and research into practice* (pp. 205–221). Mahwah, NJ: Erlbaum.
- Sternberg, R. J., Bonney, C. R., Gabora, L., Karelitz, T., & Coffin, L. (2010). Broadening the spectrum of undergraduate admissions. *College and University*, 86(1), 2–17.
- Sternberg, R. J., Castejón, J. L., Prieto, M. D., Hautamäki, J., & Grigorenko, E. L. (2001). Confirmatory factor analysis of the Sternberg Triarchic Abilities test in three international samples: An empirical test of the triarchic theory of intelligence. *European Journal of Psychological Assessment*, 17, 1–16. doi:10.1027//1015-5759.17.1.1
- Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J., Snook, S., Williams, W. M., . . . Grigorenko, E. L. (2000). *Practical intelligence in everyday life*. New York, NY: Cambridge University Press.
- Sternberg, R. J., & Grigorenko, E. L. (2000). *Teaching for successful intelligence*. Chicago, IL: Skylight.
- Sternberg, R. J., Grigorenko, E. L., Ferrari, M., & Clinkenbeard, P. A. (1999). Triarchic analysis of an aptitude–treatment interaction. *European Journal of Psychological Assessment*, 15, 3–13. doi:10.1027//1015-5759.15.1.3
- Sternberg, R. J., Grigorenko, E. L., & Jarvin, L. (2007). *Teaching for successful intelligence* (2nd ed.). Thousand Oaks, CA: Corwin Press.
- Sternberg, R. J., Grigorenko, E. L., & Zhang, L. (2008a). A reply to two stylish critiques. *Perspectives on Psychological Science*, 3, 516–517. doi:10.1111/j.1745-6924.2008.00092.x
- Sternberg, R. J., Grigorenko, E. L., & Zhang, L. (2008b). Styles of learning and thinking matter in instruction and assessment. *Perspectives on*



- Psychological Science*, 3, 486–506. doi:10.1111/j.1745-6924.2008.00095.x
- Sternberg, R. J., Jarvin, L., & Grigorenko, E. L. (2009). *Teaching for wisdom, creativity, and success*. Thousand Oaks, CA: Corwin.
- Sternberg, R. J., Jarvin, L., & Grigorenko, E. L. (2011). *Explorations of the nature of giftedness*. New York, NY: Cambridge University Press.
- Sternberg, R. J., & Lubart, T. I. (1995). *Defying the crowd*. New York, NY: Free Press.
- Sternberg, R. J., & The Rainbow Project Collaborators. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence*, 34, 321–350. doi:10.1016/j.intell.2006.01.002
- Sternberg, R. J., Torff, B., & Grigorenko, E. L. (1998). Teaching triarchically improves school achievement. *Journal of Educational Psychology*, 90, 374–384. doi:10.1037/0022-0663.90.3.374
- Sternberg, R. J., & Williams, W. M. (1996). *How to develop student creativity*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Woolfolk, A. H. (2009). *Educational psychology* (11th ed.). Upper Saddle River, NJ: Prentice-Hall.

## Appendix

### Technical Issues in the Handling of Data

This Appendix describes the technical details regarding the handling of data. Participating teachers were instructed to label and package all student materials in a particular way and send the materials to Yale. A set of materials from one test (pre- or post-) from one teacher from one unit was called a “package.” For a package to be processed and entered into the database, the student workbook had to meet the fidelity standards (see above). In collaboration with the Yale University Social Science Statistical Laboratory, an ACCESS database template was developed. This template was used to build separate databases for each of the 4 years of data collection. Each database was used to (a) inventory, or log, the materials received from the teachers; (b) track the materials as they were sent to coders to score; and (c) store test data and demographic information. The four databases were housed on the central PACE server, with file access restricted to members of the project team.

#### Database Structure

The structure of the database was rather complex and contained several types of tables, as described below.

**Participant information tables.** Four tables contained non-test information about the different participants in the study: students, teachers, schools, and districts. Unique ID numbers were given to each element within a table (e.g., each packet was given a unique ID in the packet table). Each teacher was given a different teacher ID number for each school at which he or she taught during that year; hence, some teachers were given more than one unique ID. In most cases, the information in these tables was entered into the database before any assessment data were collected.

**Coding administration tables.** Tracking test materials was a particularly challenging part of administering a large-scale project that involved continuous receipt and scoring of tests. Two important database processes were involved. The first process, material logging, was used to inventory the completed assessments received by the PACE Center. The second process, material checkout, was

used to track the assessments as they were given to coders to score. Four Access tables were involved, and information was continually added to these tables as part of the material logging and checkout processes. First, upon receiving student materials from the schools and/or teachers, PACE research assistants “logged in” the materials to the Access database for the appropriate year of data collection. The logging process consisted of assigning tests of similar type (e.g., Geometry pretests) from a single teacher to a “packet” and creating inventories of materials received by the PACE Center.

A packet was considered both (a) an envelope containing a collection of one particular test type for all the students associated with one teacher and (b) an Access database unit that identified this collection of student tests. A system of Access forms was developed to allow a research assistant simultaneously to add information to two tables that inventoried and tracked the packets and tests. The first table was a “material” table used as an inventory of test materials received by the center. The second table was a “packet” table that was used to assign a packet number to a packet and track it as it was sent to coders to score. A packet that was successfully logged in was then ready to be rated by a coder. Second, via a system of queries, packets were selected and assigned to coders to rate. A “checkout” table tracked when each coder checked out and returned each packet. In addition, a “coder” table was maintained that contained a list of each coder, his or her unique coder ID, and notes about the coder. Queries and forms were used simultaneously to update these tables as coders completed the agreement process and as packets were assigned to coders.

**Data tables.** When a coder was assigned a packet to score, she or he was also given a scoring template for the packet: an Excel spreadsheet used to record multiple-choice item responses and open-ended item ratings for each student. These templates contained a row for each student, with columns corresponding to the ratings needed for each item and additional

(Appendix continues)

columns containing identifying information (IDs for the packet, student, type of test, and coder). Each scoring template contained columns for only one test type (e.g., Geometry pretests) that corresponded to the columns of an Access data table; data for each test type were stored in separate Access data tables. The coder returned an electronic (Excel) and/or a paper copy of the completed scoring template to the National Science Foundation (NSF) team. If the coder submitted only a paper copy, it was given to data entry personnel to enter into the Excel template (with this latter option reserved for skilled coders who had little computer access or expertise). Information from the completed Excel scoring template was then directly uploaded into an Access data table by copying the data cells of the Excel sheet and pasting them into the Access table.

**Quality control.** Measures were taken to monitor the quality of the ratings during data collection. These measures included limiting database access to a small number of the most experienced personnel, using data-validation controls to prevent the entry of out-of-range values, supervising the coders carefully after their training was completed, and maintaining problem logs in the database.

**Limiting the number of Access users.** Access to the database was limited to only a small core of management personnel to ensure participant confidentiality and to minimize the possibility of human error. The databases were stored on a central server and required network permissions to be viewed or modified. For most of the study, only a small number of our most technologically sophisticated personnel were allowed access to the database to check, upload, and clean data. At times, the number of people working simultaneously on coding exceeded 20 trained coders. Rather than having all of these coders enter their ratings into the Access data tables directly, we introduced a middle step between rating and Access data entry. Coders' ratings were entered into Excel, as described above, and then given to the core database managers to upload.

**Excel template and Access table validations.** Two related measures were taken to prevent the entry of out-of-range values into the Access databases. First, cell validations were used in Excel that would allow coders to enter only legitimate ratings. Legitimate ratings included codes used to designate an omitted or illegible response to an item (i.e., 6 or e for omitted responses, and 7 or f for illegible responses). A second layer of protection was also used to prevent the uploading of empty (unrated or unrecorded) data cells into the Access database and to serve as a second check for out-of-range values. Validation rules were eventually implemented in all Access data tables to prevent the uploading of missing or out-of-range values. When a core NSF research assistant could not upload the data from a coder's template because of an out-of-range or missing value, the paper copy of the template and/or the coder was consulted to find the true value of that rating.

**Coder supervision.** Coders were not permitted to score tests until they reached an acceptable level of initial interrater reliability with their coding partner (i.e., the correlations between the pair's open-ended item ratings were greater than 0.70). Coders who

reached this criterion then began coding tests independently from their partners; coders who were not able to establish acceptable levels of interrater reliability were not permitted to continue on the project. Of the 90 coders who began the training and agreement process, only 76 were permitted to score tests for the study. Core personnel maintained weekly contact with active coders after initial interrater reliability was reached. They maintained the quality of the ratings by being available to answer coders' questions about scoring and by reminding coders of the scoring guidelines when the coder's ratings were discovered to have violated validation rules. The design of the study also allowed for the discovery of coder irregularities throughout the scoring process. As a quality control check, over thirty percent of the tests each coder scored were also scored by another coder. These overlapping ratings were used to detect discrepancies between coders and to flag coders who were having particular difficulty. Data from two of the 76 coders were deleted due to continued discrepancies with other coders. In addition, during the final stage of data cleaning before analysis, a random 10% of codings were spot-checked against the hard copies of the assessments. The 74 remaining coders were diverse with respect to their genders, ages, educational backgrounds, and test coding experience. They ranged in age from 18 to 66 and included research assistants, undergraduate student workers, temporary part-time employees, and PhD-level research scientists. Many coders had previous experience scoring tests, and some had experience creating scoring rubrics for tests. More than 20,000 pre- and posttests including over 400,000 items were read and rated by these raters. All pre- and posttests packets had two raters, who used written rubrics to evaluate the quality of children's responses. Each pair trained together on one packet: The two raters in the pair rated identical tests, their scores were then compared to establish interrater reliability, differences in scoring were pointed out and the rater pair discussed responses until agreement was reached. They were then sent back with the packet and their new ratings checked for reliability. Training was conducted until pairs reached an agreement of .70, which was treated as the minimum acceptable level; when the agreement was reached, the raters read and rated a common overlapping set of materials (representing 1/3 of all the tests rated by the pair) and then each rater read and rated separate sets. The quality of the data and interrater agreement was monitored in an ongoing fashion. Rater biases were carefully evaluated.

**Problem note tables.** During the last year of data collection, tables were created in Access to centralize notes made on test administration quirks (e.g., a teacher photocopying all but a page of a test). One table was used to record problems with a particular student; another table was used to record quirks that affected the entire classroom. Both tables made note of what was done to correct the problem. These notes were used to ensure that common problems were treated consistently. These tables were used to determine the usability of the data for final analyses.

Received July 27, 2011

Revision received November 4, 2013

Accepted December 29, 2013 ■



The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

**Manuscript preparation.** Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (6th ed.). Manuscripts may be copyedited for bias-free language (see pp. 70–77 of the *Publication Manual*). Formatting instructions (all copy must be double-spaced) and instructions on the preparation of tables, figures, references, metrics, and abstracts appear in the *Manual*. For APA's Checklist for Manuscript Submission, see [www.apa.org/pubs/journals/edu](http://www.apa.org/pubs/journals/edu). **Abstract and keywords.** All manuscripts must include an abstract containing a maximum of 250 words typed on a separate page. After the abstract, please supply up to five keywords or brief phrases. **References.** References should be listed in alphabetical order. Each listed reference should be cited in text, and each text citation should be listed in the References. Basic formats are as follows:

Hughes, G., Desantis, A., & Waszak, F. (2013). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychological Bulletin*, 139, 133–151. <http://dx.doi.org/10.1037/a0028566>

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Gill, M. J., & Sypher, B. D. (2009). Workplace incivility and organizational trust. In P. Lutgen-Sandvik & B. D. Sypher (Eds.), *Destructive organizational communication: Processes, consequences, and constructive ways of organizing* (pp. 53–73). New York, NY: Taylor & Francis.

Adequate description of participants is critical to the science and practice of educational psychology; this allows readers to assess the results, determine generalizability of findings, and make comparisons in replications, extensions, literature reviews, or secondary data analyses. Authors should see guidelines for sample–subject description in the *Manual*. Appropriate indexes of effect size or strength of relationship should be incorporated in the results section of the manuscript (see p. 34 of the *Manual*). Information that allows the reader to assess not only the significance but also the magnitude of the observed effects or relationships clarifies the importance of the findings. **Figures.** Graphics files are welcome if supplied in TIFF or EPS format. APA's policy on publication of color figures is available at <http://www.apa.org/pubs/authors/instructions.aspx?item=6>.

**Publication policies.** APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. APA policy regarding posting articles on the Internet may be found at [www.apa.org/pubs/authors/posting.aspx](http://www.apa.org/pubs/authors/posting.aspx). In addition, it is a violation of APA Ethical Principles to publish “as original data, data that have been previously published” (Standard 8.13). As this is a primary journal that publishes original material only, APA policy prohibits publication of any manuscript or data that have already been published in

whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data on which their article depends. In addition, APA Ethical Principles specify that “after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release” (Standard 8.14). Authors must have available their data throughout the editorial review process and for at least 5 years after the date of publication.

**Masked review policy.** The *Journal* has a masked review policy, which means that the identities of both authors and reviewers are masked. Every effort should be made by the authors to see that the manuscript itself contains no clues to their identities. Authors should never use first person (*I, my, we, our*) when referring to a study conducted by the author(s) or when doing so reveals the authors' identities, e.g., “in our previous work, Johnson et al., 1998 reported that . . .” Instead, references to the authors' work should be in third person, e.g., “Johnson et al. (1998) reported that . . .” The authors' institutional affiliations should also be masked in the manuscript. Authors submitting manuscripts are required to include in the cover letter the title of the manuscript along with all authors' names and institutional affiliations. However, the first page of the manuscript should omit the authors' names and affiliations, but should include the title of the manuscript and the date it is submitted. Responsibility for masking the manuscript rests with the authors; manuscripts will be returned to the author if not appropriately masked. If the manuscript is accepted, authors will be asked to make changes in wording so that the paper is no longer masked. Authors are required to state in writing that they have complied with APA ethical standards in the treatment of their sample, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained at [www.apa.org/ethics/](http://www.apa.org/ethics/) or by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242. APA requires authors to reveal any possible conflict of interest in the conduct and reporting of research (e.g., financial interests in a test procedure, funding by pharmaceutical companies for drug research). Authors of accepted manuscripts will be required to transfer copyright to APA.

**Permissions.** Authors of accepted papers must obtain and provide to the editor on final acceptance all necessary permissions to reproduce in print and electronic form any copyrighted work, including test materials (or portions thereof), photographs, and other graphic images (including those used as stimuli in experiments). On advice of counsel, APA may decline to publish any image whose copyright status is unknown.

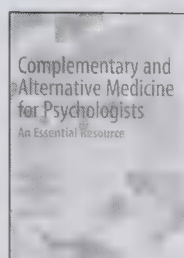
**Supplemental materials.** APA can place supplementary materials online, which will be available via the published article in the PsycARTICLES database. To submit such materials, please see [www.apa.org/pubs/authors/supp-material.aspx](http://www.apa.org/pubs/authors/supp-material.aspx) for details. Authors of accepted papers will be asked to work with the editor and production staff to provide supplementary materials as appropriate.

**Submission.** Authors should submit their manuscripts electronically via the Manuscript Submission Portal at [www.apa.org/pubs/journals/edu/index.aspx](http://www.apa.org/pubs/journals/edu/index.aspx) (follow the link for submission under Instructions to Authors). General correspondence may be addressed to the incoming editorial office at [AConley@apa.org](mailto:AConley@apa.org).



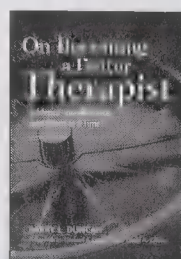
# NEW RELEASES

from the American Psychological Association



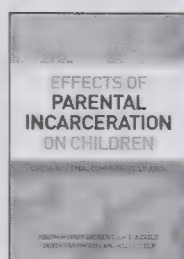
**Complementary and Alternative Medicine for Psychologists**  
*An Essential Resource*  
 Jeffrey E. Barnett, Allison Shale,  
 Gary R. Elkins, and William Ira Fisher  
 2014. 320 pages. Hardcover.

List: \$79.95 | APA Member/Affiliate: \$49.95  
 ISBN 978-1-4338-1749-6 | Item # 4317345



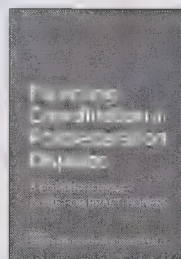
**On Becoming a Better Therapist**  
*Evidence-Based Practice One Client at a Time*  
 SECOND EDITION  
 Barry L. Duncan  
 2014. 272 pages. Hardcover.

List: \$59.95 | APA Member/Affiliate: \$49.95  
 ISBN 978-1-4338-1745-8 | Item # 4317334



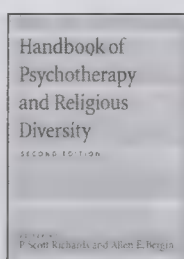
**Effects of Parental Incarceration on Children**  
*Cross-National Comparative Studies*  
 Joseph Murray, David Farrington,  
 Catrien C.J.H. Bijleveld, and Rolf Loeber  
 2014. 264 pages. Hardcover.

List: \$79.95 | APA Member/Affiliate: \$49.95  
 ISBN 978-1-4338-1743-4 | Item # 4318126



**Parenting Coordination in Postseparation Disputes**  
*A Comprehensive Guide for Practitioners*  
 Edited by Shirley Ann Higuchi  
 and Stephen J. Lally  
 2014. 264 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95  
 ISBN 978-1-4338-1739-7 | Item # 4317343



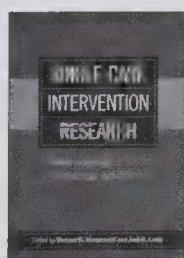
**Handbook of Psychotherapy and Religious Diversity**  
 SECOND EDITION  
 Edited by P. Scott Richards  
 and Allen E. Bergin  
 2014. 488 pages. Hardcover.

List: \$89.95 | APA Member/Affiliate: \$59.95  
 ISBN 978-1-4338-1735-9 | Item # 4317338



**Longitudinal Data Analysis Using Structural Equation Models**  
 Jack J. McArdle and  
 John R. Nesselrode  
 2014. 424 pages. Hardcover.

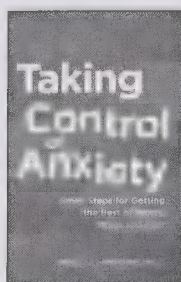
List: \$89.95 | APA Member/Affiliate: \$59.95  
 ISBN 978-1-4338-1715-1 | Item # 4316161



**Single-Case Intervention Research**  
*Methodological and Statistical Advances*  
 Edited by Thomas R. Kratochwill  
 and Joel R. Levin  
 2014. 408 pages. Hardcover.

■ Series: Division 16:  
 School Psychology

List: \$79.95 | APA Member/Affiliate: \$49.95  
 ISBN 978-1-4338-1751-9 | Item # 4316163



AN APA LIFETOOLS® BOOK  
**Taking Control of Anxiety**  
*Small Steps for Getting the Best of Worry, Stress, and Fear*  
 Bret A. Moore, PsyD  
 2014. 248 pages. Paperback.

List: \$16.95 | APA Member/Affiliate: \$16.95  
 ISBN 978-1-4338-1747-2 | Item # 4441023



AMERICAN PSYCHOLOGICAL ASSOCIATION

TO ORDER: 800-374-2721 • [www.apa.org/pubs/books](http://www.apa.org/pubs/books)

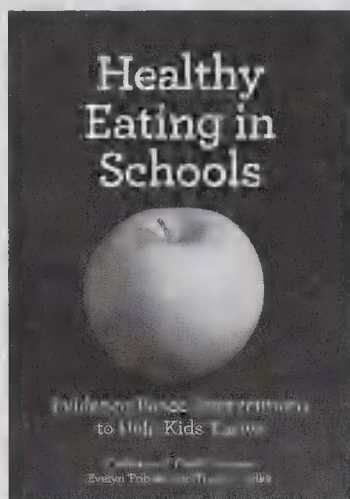
APA/PA



# HEALTHY EATING IN SCHOOLS

## Evidence-Based Interventions to Help Kids Thrive

Catherine P. Cook-Cottone, Evelyn Tribole, and Tracy L. Tylka



Concern over increased childhood obesity has spurred various school-based interventions. However, these interventions often have little positive effect and may inadvertently contribute to unhealthy behaviors in attempt to lose weight. Indeed, a general emphasis on appearance and weight (rather than health) can promote eating disordered behaviors.

This book provides a conceptual model for understanding both obesity and eating disordered behaviors. Specifically, it advocates for body acceptance and intuitive eating—a flexible, healthy eating behavior involving awareness of the body's hunger and satiety cues. Within this context, the chapters review evidence-based school interventions in nutrition, self-regulation, exercise, body acceptance, media literacy, and mindfulness. Guidance is also provided for identifying, referring, and supporting students with emerging eating disorders.

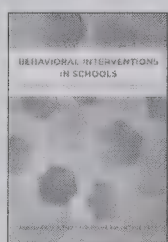
Without empirically supported guidance, schools run the risk of implementing ineffective or harmful programming in an effort to do good. Thus, this book is a much needed resource for teachers, administrators, counselors, nurses, and other school personnel. **Series: Division 16: School Psychology. 2013. 288 pages. Hardcover.**

List: \$69.95 | APA Member/Affiliate: \$49.95 | ISBN 978-1-4338-1300-9 | Item # 4317303

### CONTENTS:

**I. Conceptual Framework** | 1. Defining Healthy and Intuitive Eating | 2. Why We Eat the Way We Do: The Role of Personal and External Factors | **II. The Healthy Student Approach** | 3. Rationale for a Three-Pillar Approach | 4. Pillar I: Intuitive Eating and Nutrition | 5. Pillar II: Healthy Physical Activity | 6. Pillar III: Mindfulness, Self-Care, and Emotional Regulation | **III. School-Based Interventions and Policies** | 7. Preventative Intervention: Media Literacy, Body Image, Body Tolerance, and Self-Regulated Eating | 8. Screening, Assessing, and Supporting Students with Eating and Body Image Problems | 9. Federal School Food Policies and Professional Guidelines | Appendix A: Definitions of Uncommon Disorders of Eating | Appendix B: Children's Eating Attitudes Test | Appendix C: Intuitive Eating Scale for Adolescents | Appendix D: Body Appreciation Scale | Appendix E: Sociocultural Attitudes Towards Appearance Questionnaire-3 (SATAQ-3; Adolescent Version) | Suggested Resources for School Personnel on Healthy Eating

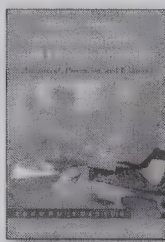
### ALSO OF INTEREST



AVAILABLE ON  
AMAZON KINDLE®  
**Behavioral  
Interventions  
in Schools  
Evidence-Based  
Positive Strategies**  
Edited by

Angeleque Akin-Little, Steven  
G. Little, Melissa A. Bray, and  
Thomas J. Kahle  
2009. 350 pages. Hardcover.

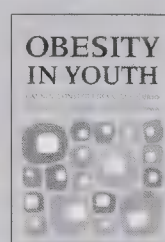
List: \$49.95 | APA Member/Affiliate: \$39.95  
ISBN 978-1-4338-0460-1 | Item # 4317189  
CEP Credits: 8



AVAILABLE ON  
AMAZON KINDLE®  
**Body Image,  
Eating  
Disorders, and  
Obesity in Youth  
Assessment,  
Prevention,**

**and Treatment**  
**SECOND EDITION**  
Edited by Linda Smolak and  
J. Kevin Thompson  
2009. 389 pages. Hardcover.

List: \$29.95 | APA Member/Affiliate: \$24.95  
ISBN 978-1-4338-0405-2 | Item # 4317167  
CEP Credits: 9



AVAILABLE ON  
AMAZON KINDLE®  
**Obesity  
in Youth  
Causes,  
Consequences,  
and Cures**

Edited by Leslie J. Heinberg  
and J. Kevin Thompson  
2009. 243 pages. Hardcover.

List: \$49.95 | APA Member/Affiliate: \$39.95  
ISBN 978-1-4338-0427-4 | Item # 4317176  
CEP Credits: 4



AMERICAN PSYCHOLOGICAL ASSOCIATION

**APA BOOKS ORDERING INFORMATION: 800-374-2721 • [www.apa.org/pubs/books](http://www.apa.org/pubs/books)**

In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502

In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972

AD2651



# BEST SELLERS

from the American Psychological Association



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

## Assessing Bilingual Children in Context

An Integrated Approach

Edited by Amanda B. Clinton

2014. 281 pages. Hardcover.

**Series: Division 16: School Psychology**

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1565-2 | Item # 4317323

## Universal Screening in Educational Settings

Evidence-Based Decision Making for Schools

Edited by Ryan J. Kettler, Todd A. Glover,

Craig A. Albers, Kelly A. Feeney-Kettler

2014. 328 pages. Hardcover.

**Series: Division 16: School Psychology**

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1550-8 | Item # 4317318

## Educational Evaluations of Children With Special Needs

Clinical and Forensic Considerations

David Breiger, Kristen Bishop, and G.

Andrew H. Benjamin

2014. 152 pages. Hardcover.

List: \$59.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1575-1 | Item # 4317326

## Autism Spectrum Disorder

A Clinical Guide for General Practitioners

V. Mark Durand

2014. 216 pages. Hardcover.

List: \$59.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1569-0 | Item # 4317325

## The Stigma of Disease and Disability

Understanding Causes

and Overcoming Injustices

Edited by Patrick W. Corrigan

2014. 312 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1583-6 | Item # 4318124

## Attachment-Based Family Therapy for Depressed Adolescents

Guy S. Diamond, Gary M. Diamond,

and Suzanne A. Levy

2014. 280 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1567-6 | Item # 4317324

## Mechanisms of Social Connection

From Brain to Group

Edited by Mario Mikulincer

and Phillip R. Shaver

2014. 416 pages. Hardcover.

List: \$79.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1415-0 | Item # 4318118

## Varieties of Anomalous Experience

Examining the Scientific Evidence

SECOND EDITION

Edited by Etzel Cardena, Steven Jay Lynn,

and Stanley Krippner

2014. 464 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1529-4 | Item # 4316157

## APA Handbook of Multicultural Psychology

Volume 1: Theory and Research

Volume 2: Applications and Training

Editor-in-Chief Frederick T. L. Leong

2014. 1,314 pages. Hardcover.

List: \$395.00 | APA Member/Affiliate: \$195.00

ISBN 978-1-4338-1255-2 | Item # 4311511

## APA Dictionary of Statistics and Research Methods

Editor-in-Chief Sheldon Zedeck

2014. 452 pages. Hardcover.

List: \$39.95 | APA Member/Affiliate: \$29.95

ISBN 978-1-4338-1533-1 | Item # 4311019

## APA Handbook of Sexuality and Psychology

Volume 1: Person-Based Approaches

Volume 2: Contextual Approaches

Editors-in-Chief Deborah L. Tolman

and Lisa M. Diamond

2014. 1,400 pages. Hardcover.

List: \$395.00 | APA Member/Affiliate: \$195.00

ISBN 978-1-4338-1369-6 | Item # 4311512

## Psychotherapy Theories and Techniques

A Reader

Edited by Gary R. VandenBos, Edward

Meidenbauer, and Julia Frank-McNeil

2014. 368 pages. Paperback.

List: \$34.95 | APA Member/Affiliate: \$29.95

ISBN 978-1-4338-1619-2 | Item # 4317329

## Exploring Three Approaches to Psychotherapy

Leah S. Greenberg, Nancy McWilliams,

Amy Wenzel

2014. 280 pages. Hardcover.

List: \$39.95 | APA Member/Affiliate: \$34.95

ISBN 978-1-4338-1521-8 | Item # 4317316

Paperback:

List: \$29.95 | APA Member/Affiliate: \$24.95

ISBN 978-1-4338-1520-1 | Item # 4317315

## Exploring Sport and Exercise Psychology

THIRD EDITION

Edited by Judy L. Van Raalte

and Britton W. Brewer

2014. 672 pages. Paperback.

List: \$49.95 | APA Member/Affiliate: \$39.95

ISBN 978-1-4338-1357-3 | Item # 4317312

## The Power of Metaphor

Examining Its Influence on Social Life

Edited by Mark J. Landau, Michael D.

Robinson, and Brian P. Meier

2014. 304 pages. Hardcover.

List: \$79.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1579-9 | Item # 4318123

## Culture Reexamined

Broadening Our Understanding of Social

and Evolutionary Influences

Edited by Adam B. Cohen

2014. 256 pages. Hardcover.

List: \$79.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1587-4 | Item # 4316159

## Medical Family Therapy and Integrated Care

SECOND EDITION

Susan H. McDaniel, William J. Doherty,

and Jeri Hepworth

2014. 368 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1518-8 | Item # 4317314

## The Marriage Checkup Practitioner's Guide

Promoting Lifelong Relationship Health

James V. Cordova

2014. 264 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1552-2 | Item # 4317319

Available on Amazon Kindle®

## The Nature of Work

Advances in Psychological Theory,

Methods, and Practice

Edited by J. Kevin Ford,

John R. Hollenbeck, and Ann Marie Ryan

2014. 328 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1537-9 | Item # 4318119

## Couple and Family Therapy

An Integrative Map of the Territory

Jay L. Lebow

2014. 312 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1362-7 | Item # 4317313

## Treatment Integrity

A Foundation for Evidence-Based

Practice in Applied Psychology

Edited by Lisa M. Hagermoser Sanetti

and Thomas R. Kratochwill

2014. 320 pages. Hardcover.

**Series: Division 16: School Psychology**

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1581-2 | Item # 4317327

## Pretend Play in Childhood

Foundation of Adult Creativity

Sandra W. Russ

2014. 240 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1561-4 | Item # 4318122

## Trauma and Substance Abuse

Causes, Consequences, and

Treatment of Comorbid Disorders

SECOND EDITION

Edited by Paige Quimette

and Jennifer P. Read

2014. 336 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-4338-1523-2 | Item # 4317317

## Geographical Psychology

Exploring the Interaction of

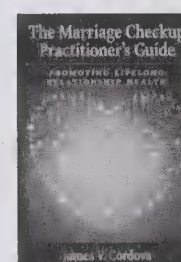
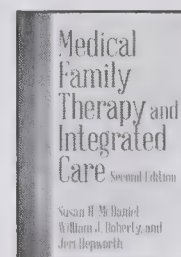
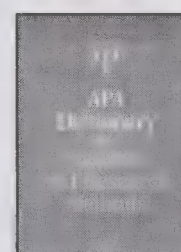
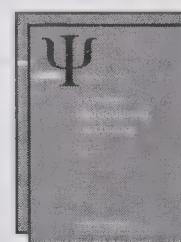
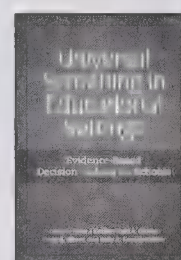
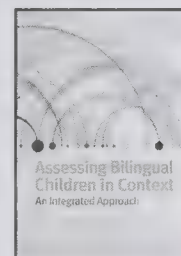
Environment and Behavior

Edited by Peter J. Rentfrow

2014. 336 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95

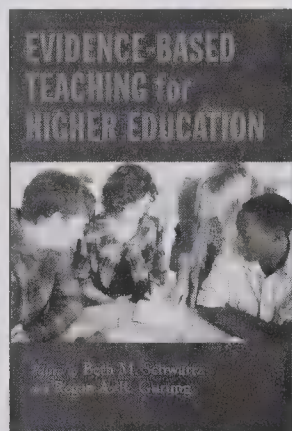
ISBN 978-1-4338-1539-3 | Item # 4316158





# EVIDENCE-BASED TEACHING FOR HIGHER EDUCATION

Edited by Beth M. Schwartz and Regan A. R. Gurung



Over the past two decades, a growing body of scholarship of teaching and learning (SoTL) has emerged. This empirical study of teaching methods, course design, and students' study practices has yielded invaluable information about how teachers teach and learners learn. Yet, university faculty members remain largely unaware of the findings of SoTL research. As a result, they tend to choose their teaching techniques and tools based on intuition and previous experience rather than on scientific evidence of effectiveness.

This book synthesizes SoTL findings to help teachers choose techniques and tools that maximize student learning. Evidence-based recommendations are provided regarding teacher-student rapport, online teaching, use of technology in the classroom (such as audience response systems, podcasting, blogs, and wikis), experiential learning (such as internships, teaching assistantships, research assistantships, and in-class research projects), students' study habits, and more.

In order to stimulate future SoTL research, the book also recommends numerous areas for future investigation. It concludes with advice for documenting teaching effectiveness for tenure review committees.

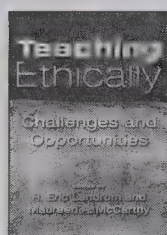
Both novice and experienced university teachers will find this book useful, as well as professionals who work in faculty development centers. 2012. 168 pages. Paperback.

List: \$39.95 | APA Member/Affiliate: \$39.95 | ISBN 978-1-4338-1172-2 | Item # 4317288

## CONTENTS

Foreword, William Buskist | Acknowledgments | Introduction, Beth M. Schwartz and Regan A. R. Gurung | 1. Benefits of Using SoTL in Picking and Choosing Pedagogy, Randolph A. Smith | 2. Building Rapport in the Classroom and Student Outcomes, Janie H. Wilson, Shauna B. Wilson, and Angela M. Legg | 3. Using Technology to Enhance Teaching and Learning, Christopher R. Poirier and Robert S. Feldman | 4. Online Teaching, Chandra M. Mehrotra and Lawrence McGahey | 5. Experiential Learning, Kristin M. Vespia, Georjeanna Wilson-Doenges, Ryan C. Martin, and Deirdre M. Radosovich | 6. How Should Students Study?, Regan A. R. Gurung and Lee I. McCann | 7. Selection of Textbooks or Readings for Your Course, R. Eric Landrum | 8. Are You Really Above Average? Documenting Your Teaching Effectiveness, Jane S. Halonen, Dana S. Dunn, Maureen A. McCarthy, and Suzanne C. Baker

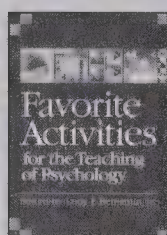
## ALSO OF INTEREST



**Teaching Ethically: Challenges and Opportunities**  
Edited by  
R. Eric Landrum  
and

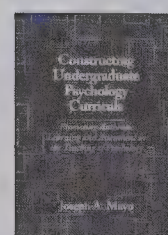
Maureen A. McCarthy  
2012. 214 pages. Hardcover.

List: \$49.95 | APA Member/Affiliate: \$39.95  
ISBN 978-1-4338-1086-2 | Item # 4311035



**Favorite Activities for the Teaching of Psychology**  
Edited by  
Ludy T. Benjamin, Jr.  
2008. 291 pages.  
Paperback.

List: \$34.95 | APA Member/Affiliate: \$29.95  
ISBN 978-1-4338-0349-9 | Item # 4316105



**Constructing Undergraduate Psychology Curricula**  
Promoting Authentic Learning and Assessment in  
the Teaching of Psychology  
Joseph A. Mayo

2010. 227 pages. Hardcover.

List: \$39.95 | APA Member/Affiliate: \$29.95  
ISBN 978-1-4338-0563-9 | Item # 4316116



AMERICAN PSYCHOLOGICAL ASSOCIATION

**APA BOOKS ORDERING INFORMATION: 800-374-2721 • [www.apa.org/pubs/books](http://www.apa.org/pubs/books)**

In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502

In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972

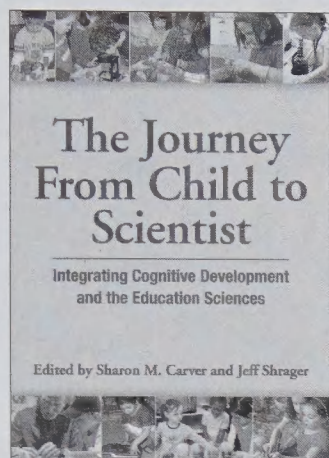
AD2616



# THE JOURNEY FROM CHILD TO SCIENTIST

## *Integrating Cognitive Development and the Education Sciences*

Edited by Sharon M. Carver and Jeff Shrager



The impulse to investigate the natural world is deeply rooted in our earliest childhood experiences. This notion has long guided researchers to uncover the cognitive mechanisms underlying the development of scientific reasoning in children.

Until recently, however, research in cognitive development and education followed largely independent tracks. A major exception to this trend is represented in the multifaceted work of David Klahr. His lifelong effort to integrate a detailed understanding of children's reasoning and skill acquisition with the role of education in influencing and facilitating scientific exploration has been essential to the growth of these fields.

In this volume, a diverse group of stellar contributors follow Klahr's example in examining the practical implications of our insights into cognitive development for children in the classroom. Authors discuss such wide-ranging ideas as the evolution of "folk science" in young children and the mechanisms that underlie mathematical understanding, as well as mental models used by children in classroom activities.

The volume's lessons will have profound implications for STEM education, and for the next generation of scientists.

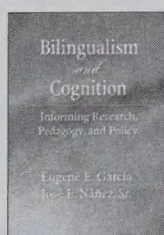
**Series: Decade of Behavior.** 2012. 352 pages. Hardcover.

List: \$59.95 | APA Member/Affiliate: \$49.95 | ISBN 978-1-4338-1138-8 | Item # 4318104

### Contents

**Introduction:** A Life's Journey From Child to Scientist, Sharon M. Carver and Jeff Shrager | **1.** From Theory to Application and Back: Following in the Giant Footsteps of David Klahr, Robert Siegler | **2.** The Learning of Science and the Science of Learning: The Role of Analogy, Zhe Chen | **3.** Does Folk Science Develop?, Frank Keil | **4.** The Evolved Mind and Scientific Discovery, David C. Geary | **5.** Applying the Klahrian Method Toward an Integrated Discipline of Educational Neuroscience, Kevin Dunbar | **6.** Is Development Domain Specific or Domain General? A Third Alternative, Annette Karmiloff-Smith | **7.** Simulating Discovery and Education in a Soccer Science World, Jeff Shrager | **8.** Moving Young Scientists in Waiting Onto Science Learning Pathways: Focus on Observation, Rachel Gelman and Kimberly Brenneman | **9.** Supporting Inquiry About the Foundations of Evolutionary Thinking in the Elementary Grades, Richard Lehrer and Leona Schauble | **10.** Engineering in and for Science Education, Christian D. Schunn, Eli M. Silk, and Xornam S. Apedoe | **11.** To Teach or Not to Teach Through Inquiry, Erin Marie Furtak, Richard J. Shavelson, Jonathan T. Shemwell, and Maria Figueroa | **12.** Epistemic Foundations for Conceptual Change, Richard A. Duschl and Maria Pilar Jimenez-Aleixandre | **13.** Patterns, Rules, and Discoveries in Life and in Science, David Klahr

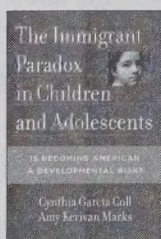
### ALSO OF INTEREST



**Bilingualism and Cognition**  
*Informing Research, Pedagogy, and Policy*  
Eugene E. García  
and  
José E. Náñez, Sr.

2012. 242 pages. Hardcover.

List: \$39.95 | APA Member/Affiliate: \$34.95  
ISBN 978-1-4338-0879-1 | Item # 4318087



AVAILABLE ON  
AMAZON KINDLE®  
**The Immigrant Paradox in Children and Adolescents**  
*Is Becoming American a Developmental Risk?*

Edited by Cynthia García Coll  
and Amy Kerivan Marks  
2012. 328 pages. Hardcover.

List: \$49.95 | APA Member/Affiliate: \$39.95  
ISBN 978-1-4338-1053-4 | Item # 4318097



**The Development of Giftedness and Talent Across the Life Span**

Edited by Frances Degen Horowitz, Rena F. Subotnik, and Dona J. Matthews  
2009. 252 pages. Hardcover.

List: \$49.95 | APA Member/Affiliate: \$39.95  
ISBN 978-1-4338-0414-4 | Item # 4318051  
CEP Credit: 5



AMERICAN PSYCHOLOGICAL ASSOCIATION

**APA BOOKS ORDERING INFORMATION: 800-374-2721 • www.apa.org/pubs/books**

In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502

In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972

AD2583




# BEHAVIORAL INTERVENTIONS IN SCHOOLS

## EVIDENCE-BASED POSITIVE STRATEGIES

Edited by Angeleque Akin-Little, Steven G. Little, Melissa A. Bray,  
and Thomas J. Kehle

AVAILABLE ON AMAZON KINDLE®

The emotional and behavioral problems of students in the classroom are a major concern for teachers, administrators, and the public. This book provides school psychologists, counselors, social workers, school administrators, and teachers with a summary of ecologically sound primary, secondary, and tertiary prevention strategies. 2009. 350 pages. Hardcover.  Credit: 8

### CONTENTS:

#### PART I. FOUNDATIONS FOR DESIGNING SCHOOL-BASED BEHAVIORAL INTER-

**VENTIONS** ■ Chapter 1. Behavioral Consultation, *William P. Erchul* and *Ann C. Schulte* ■ Chapter 2. Behavioral Assessment in the Schools, *T. Steuart Watson* and *Tonya Watson* ■ Chapter 3. Introduction to Functional Behavioral Assessment, *George H. Noell* and *Kristin A. Gansle* ■ Chapter 4. The Importance of Treatment Integrity in School-Based Behavioral Intervention, *Brian K. Martens* and *Laura Lee McIntyre* ■ Chapter 5. The True Effects of Extrinsic Reinforcement on "Intrinsic" Motivation, *Angeleque Akin-Little* and *Steven Little*

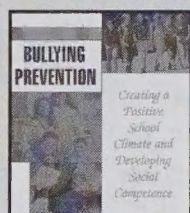
**PART II. SYSTEMATIC APPROACHES TO PREVENTION AND INTERVENTION** ■ Chapter 6. Contributions of Cognitive Behavior Therapy to School Psychology, *Raymond DiGiuseppe* ■ Chapter 7. Improving Children's Fluency in Reading, Mathematics, Spelling, and Writing: A Review of Evidence-Based Academic Interventions, *Tanya L. Eckert*, *Robin M. Coddington*, *Adrea J. Truckenmiller*, and *Jennifer L. Rheinheimer* ■ Chapter 8. School-Wide Positive Behavior Support: A Systems Level Application of Behavioral Principles, *Brandi Simonsen* and *George Sugai* ■ Chapter 9. Classroom Management, *Joseph Webby* and *Kathleen Lynne Lane* ■ Chapter 10. Applying Group-Oriented Contingencies in the Classroom, *Christopher Skinner*, *Amy L. Skinner*, and *Bobbie Burton* ■ Chapter 11. Classroom Application of Reductive Procedures: A Positive Approach, *Steven G. Little*, *Angeleque Akin-Little*, and *Clayton R. Cook* ■ Chapter 12. Generalization and Maintenance of Learned Positive Behavior, *Mark W. Steege* and *Erin Sullivan*

**PART III. SPECIFIC BEHAVIORAL TECHNIQUES** ■ Chapter 13. Using Response to Intervention for Identification of Specific Learning Disabilities, *Frank M. Gresham* ■ Chapter 14. Daily Report Cards: Home Based Consequences for Classroom Behavior, *Mary Lou Kelley* and *Nichole Jurbergs* ■ Chapter 15. Self-Modeling, *Thomas Kehle* and *Melissa A. Bray*

**PART IV. CUSTOMIZING BEHAVIORAL STRATEGIES FOR SPECIAL POPULATIONS** ■ Chapter 16. Practical Strategies in Working With Difficult Students, *William R. Jenson*, *Elaine Clark*, and *Jason Burrow-Sanchez* ■ Chapter 17. Behavioral Interventions With Externalizing Disorders, *George J. DuPaul* and *Lisa L. Weyandt* ■ Chapter 18. Interventions for Internalizing Disorders, *Thomas J. Huberty* ■ Chapter 19. Behavioral Interventions for Preschoolers, *David W. Barnett* and *Renee O. Hawkins* ■ Chapter 20. Behavioral Interventions and Autism in the Schools, *Susan M. Wilczynski*, *Laura Fisher*, *Lauren Christian*, and *Jesse Logue* ■ Chapter 21. Trauma Focused Cognitive Behavior Therapy, *Steven G. Little* and *Angeleque Akin-Little*

ISBN 978-1-4338-0460-1 • Item # 4317189 • List: \$49.95 • APA Member/Affiliate: \$39.95

### ALSO AVAILABLE

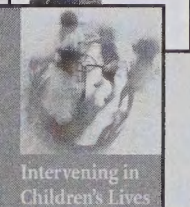


AVAILABLE ON AMAZON KINDLE®

#### BULLYING PREVENTION

**Creating a Positive School Climate and Developing Social Competence**


Pamela Orpinas and Arthur M. Horne  
2006 • 293 pages • Hardcover  
ISBN 978-1-59147-282-7 • Item # 4317082  
List: \$24.95 • APA Member/Affiliate: \$19.95



AVAILABLE ON AMAZON KINDLE®

#### INTERVENING IN CHILDREN'S LIVES

**An Ecological, Family-Centered Approach to Mental Health Care**

Thomas J. Dishion and Elizabeth A. Stormshak  
2007 • 319 pages • Hardcover  
ISBN 978-1-59147-428-9 • Item # 4317115  
List: \$19.95 • APA Member/Affiliate: \$19.95  
 Credit: 6



#### CONDUCTING SCIENCE-BASED PSYCHOLOGY RESEARCH IN SCHOOLS

Edited by Lisa M. Dinella  
2009 • 225 pages • Hardcover  
ISBN 978-1-4338-0468-7 • Item # 4317197  
List: \$29.95 • APA Member/Affiliate: \$24.95

**APA Books**  
Ordering Information  
**800-374-2721**  
[www.apa.org/pubs/books](http://www.apa.org/pubs/books)  
In Washington, DC,  
call: 202-336-5510  
TDD/TTY: 202-336-6123  
Fax: 202-336-5502  
In Europe, Africa, or the Middle East,  
call: +44 (0) 1767 604972 AD2481



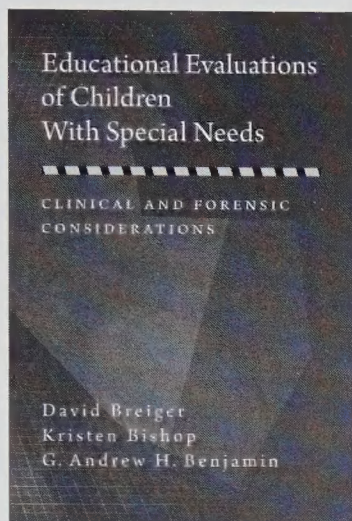
AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION



# EDUCATIONAL EVALUATIONS OF CHILDREN WITH SPECIAL NEEDS

## *Clinical and Forensic Considerations*

David Breiger, Kristen Bishop, and G. Andrew H. Benjamin



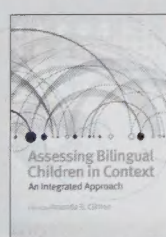
This book describes how to perform an independent educational evaluation for children with special needs. Chapters describe the suggested format and content of initial meetings with parents and school officials, the assessment and evaluation process, how to piece together the final report, and additional issues that arise after the final settlement, including testimony in due process hearings. They also carefully outline the evaluator's responsibilities under the law. Perhaps most importantly, they provide crucial suggestions for how evaluators can navigate conflict that often arises between parents and school officials, while remaining focused on providing the best possible education for all children. 2014. 152 pages. Hardcover.

.....  
List: \$59.95 | APA Member/Affiliate: \$49.95 | ISBN 978-1-4338-1575-1 | Item # 4317326

### CONTENTS:

**Introduction** | **Chapter 1.** Context and History of Special Education Evaluations | **Chapter 2.** Law, Ethics, and Competence | **Chapter 3.** Referral, Clinical Interview, and Psychological Assessment | **Chapter 4.** Concluding Evaluation and Feedback | **Chapter 5.** Final Report | **Chapter 6.** Presentation in Due Process Hearings and Postulating Interactions | **Afterword** | **Appendix A.** Sample Independent Educational Evaluation Report | **Appendix B.** Independent Educational Evaluation: Parents' Agreement | **Appendix C.** Summary of Independent Educational Evaluation for Parents | **Appendix D.** Common Mistakes to Avoid While Conducting Independent Educational Evaluations

### ALSO OF INTEREST



#### Assessing Bilingual Children in Context

*An Integrated Approach*

Edited by  
Amanda B. Clinton

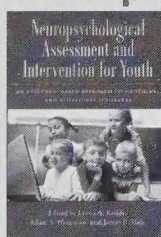
2014. 281 pages. Hardcover.

• Series: Division 16 / School Psychology

.....  
List: \$69.95 | APA Member/Affiliate: \$49.95  
ISBN 978-1-4338-1565-2 | Item # 4317323

AVAILABLE ON AMAZON KINDLE®

#### Neuropsychological Assessment and Intervention for Youth



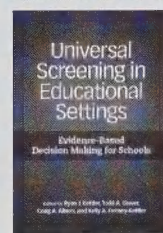
*An Evidence-Based Approach to Emotional and Behavioral Disorders*

Edited by Linda A. Reddy, Adam S. Weissman, and James B. Hale

2013. 364 pages. Hardcover.

.....  
List: \$69.95 | APA Member/Affiliate: \$49.95  
ISBN 978-1-4338-1266-8 | Item # 4316149

#### Universal Screening in Educational Settings



*Evidence-Based Decision Making for Schools*

Edited by  
Ryan J. Kettler,

Todd A. Glover, Craig A. Albers,  
and Kelly A. Feeney-Kettler

2014. 328 pages. Hardcover.

• Series: Division 16 / School Psychology

.....  
List: \$69.95 | APA Member/Affiliate: \$49.95  
ISBN 978-1-4338-1550-8 | Item # 4317318



AMERICAN PSYCHOLOGICAL ASSOCIATION

**APA BOOKS ORDERING INFORMATION: 800-374-2721 • [www.apa.org/pubs/books](http://www.apa.org/pubs/books)**

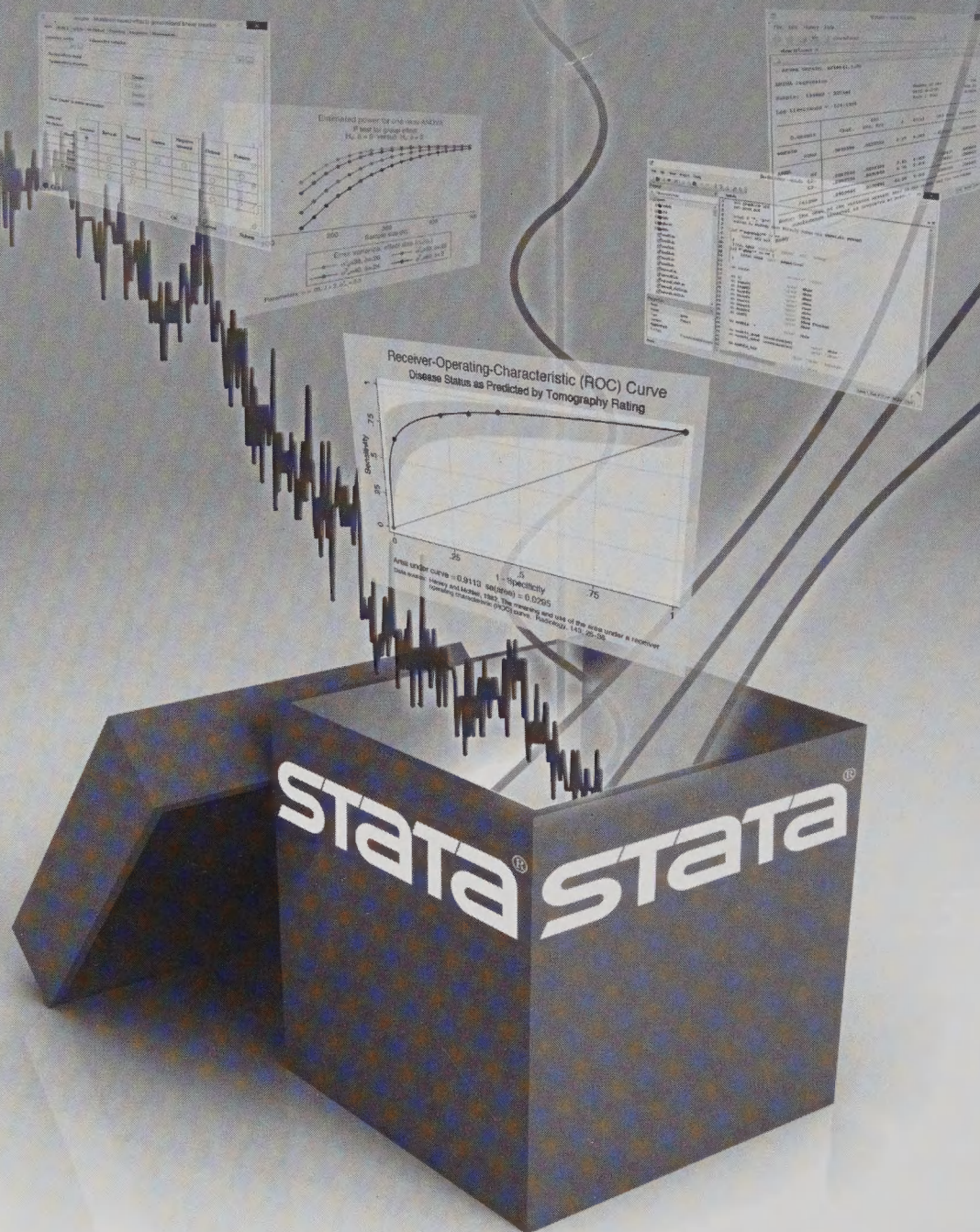
In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502

In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972

AD2342



# STATA<sup>®</sup> does more.



**Get the ease of a user-friendly interface with the flexibility of a coder's paradise.**

Stata's clean interface is arranged to simplify your workflow. The Data Editor, graph editor, and dialog boxes ease all types of analyses. But there are no restrictions. With Stata's intuitive command syntax and matrix programming language, you have the freedom to customize Stata to perfectly suit your needs.

ANOVA, CFA, hierarchical models, growth curves, interaction plots, SEM ... Stata does all this, and more.

**One unified statistical software program for all your analytical needs.**

**STATA<sup>®</sup>**



**[stata.com/edu14](http://stata.com/edu14)**